Universidade de Aveiro
Departamento de
Electrónica, Telecomunicações e Informática
2011

JAMES
MONTEIRO
APARÍCIO

A SIMETRIA DAS SEQUÊNCIAS DE ADN

DNA SYMMETRY

JAMES
MONTEIRO
APARÍCIO

A SIMETRIA DAS SEQUÊNCIAS DE ADN

DNA SYMMETRY

dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Dr. Carlos Alberto da Costa Bastos, Professor auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e da Dra. Vera Mónica Almeida Afreixo, Professora auxiliar do Departamento de Matemática da Universidade de Aveiro

o júri

presidente

Prof. Dr. Armando José Formoso de Pinho
professor associado com agregação da Universidade de Aveiro

vogais

Prof. Dr. Carlos Alberto da Costa Bastos
professor auxiliar da Universidade de Aveiro

Profa. Dra. Vera Mónica Almeida Afreixo
professora auxiliar da Universidade de Aveiro

Prof. Dr. José Paulo Ferreira Lousado
professor adjunto da Escola Superior de Tecnologia e Gestão de Lamego do Instituto Politécnico de Viseu

agradecimentos

Em primeiro lugar, quero agradecer aos meus orientadores, Dr. Carlos Alberto da Costa Bastos e Dra. Vera Mónica Almeida Afreixo, pelo apoio, empenho e disponibilidade ao longo desta Dissertação de Mestrado.
Quero ainda agradecer aos meus colegas do IEETA e por último, quero dedicar este trabalho às pessoas mais importantes da minha vida. Os meus mais sinceros agradecimentos aos meus pais, ao meu irmão, à minha cunhada e à minha namorada por todo o apoio e por acreditarem em mim.

palavras-chave simetria no ADN, nucleótides, dinucleótidos, trinucleótidos, oligonucleotidos, genome, testes de equivalência

resumo

A investigação do DNA tem sido uma das áreas de investigação mais exploradas no último século. Desde a sua primeira descrição até à primeira sequência completa do genoma humano muito foi descoberto, mas ainda estamos longe de o compreender completamente.

Neste trabalho tentámos explorar a ordem até à qual se verifica a existência de simetria relevante em genomas, e para esse fim, usámos um conjunto de genomas de vários organismos. Tentámos encontar relação entre os vários genomas através das características de simetria.

Foram analisados três tipos de simetria: simetria inversa, simetria reversa e simetria complementar. Usámos, ainda, uma nova medida para classificar a simetria: a proporção de pares equivalentes.

A natureza das operações envolvidas, o tamanho da memória e a eficiência temporal são factores a ter em conta aquando do desenvolvimento de ferramentas computacionais. Várias soluções foram exploradas tendo como objectivo minimizar a memória utilizada e minimizar o tempo de execução.

Confirma-se uma tendência para a existência de simetria inversa no conjunto dos genomas usados e observou-se que existe associação entre os resultados das medidas de simetria e o tamanho dos genomas.

keywords DNA symmetry, nucleotides, dinucleotides, trinucleotides, oligonucleotides, genome, equivalence tests

abstract DNA research has been one of the most explored areas in the last century. From its first description to the first complete human genome sequence a lot has been discovered, but we are still far from fully understanding it.
With this work we tried to find until which order is relevant symmetry found in genomes and for that purpose, we used several genomes of different organisms. We tried to find a relation between the various genomes by analysing their symmetry characteristics.
Three types of symmetry were analysed: complementary symmetry, reverse symmetry, and inverted symmetry. Also, a new symmetry measure was used: the proportion of equivalent pairs.
The nature of the operations involved, memory space and time efficiency are important factors to be considered when developing computational tools. A few different solutions are explored in order to minimize memory allocation and minimize runtimes.
This work confirms a tendency for the inverted symmetry in the set of genomes used and it was also observed an association between the symmetry measure results and the size of the genomes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

DNA (deoxyribonucleic acid) research has been one of the most explored areas in the last century. From its first description to the first complete human genome sequence a lot has been discovered, but we are still far from fully understanding it. Discovering DNA has allowed countless scientists and researchers to uncover its structure and behaviour. Leading to better understanding of diseases, detecting predisposition to them as well as contributing to the development of treatments. These breakthroughs now offer hope for patients who suffer from what were once untreatable diseases. Other fields beside medicine have benefited from the understanding of DNA, such as forensic science, agriculture or even legal issues such as paternity cases.

This dissertation will focus on trying to further understand the DNA organization, specifically the symmetry found on DNA. Several studies have been done in this field. Some with focus only on dinucleotide and trinucleotide symmetry [3, 2], others higher oligonucleotide orders [15], others on using different symmetry (inverted, reversed and complemented) and similarity measures [4] and others on high olygonucleotide orders [20]. The objective of this dissertation is to implement an extensive tool, capable of analysing DNA sequences at different oligonucleotide orders, different kinds of symmetry, using various types of symmetry and similarity measures, process a large number of different genomes and analyse the results.

In the first chapter some basic concepts of molecular biology will be introduced, Chargaff's rules, the identification of some possible types of symmetry, as well as a review of the state of the art in the field. Chapter two will focus on the process of information retrieval, followed by compression methods on chapter three. Chapter four will cover the different symmetry and similarity measures used. In chapter five we will analyse the results obtained, followed by conclusions and suggestions for possible future work in the last chapter.

1.1 Basic concepts of molecular biology

DNA Structure

Nucleic Acids were first discovered by Friedrisch Miescher in 1869 [8]. It was initially thought that these acids were only found in the nucleus, therefore its designation. Nowadays it is well known that nucleic acids can be found outside the nucleus, nevertheless its designation is still used. Deoxyribonucleic acid (DNA) is one of these nucleic acids and it is found in every cell.

DNA contains a sequence of four different nucleotides. Each nucleotide consists of a deoxyribose connected to a phosphate and a nitrogenous base (see figure 1.1). These bases are (see figure 1.2):

- Adenine(A).
- Thymine(T).
- Cytosine(C).
- Guanine(G).

Adenine and guanine are purines, being constituted by double-ringed structures, and cytosine and thymine are pyrimidines, single-ringed structures.

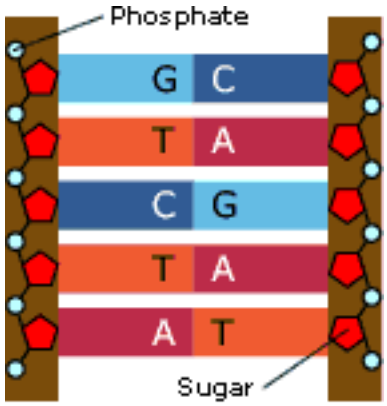

Figure 1.1: DNA basic structure (from [12]).

The DNA structure was firstly described in 1953 by Watson and Crick (using x-ray diffraction data) as a spiral staircase structure, proposing the double helix model [23] (see figure 1.3). Watson and Crick already knew that the quantity of guanine and cytosine were the same, as well as the amount of adenine and thymine from Erwin Chargaff's work [6].

According to their model, a DNA molecule consists of two nucleotide strands coiled around each other in a spiral staircase structure. The sugar and the phosphate are on the outside of

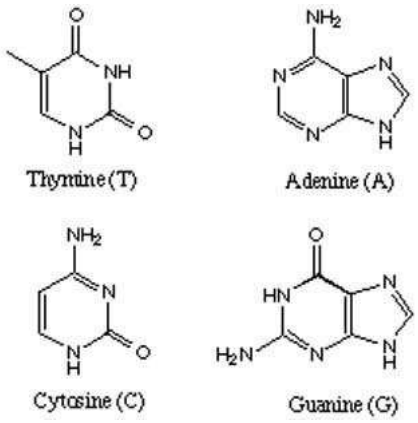

Figure 1.2: Nitrogenous bases of DNA.

the double helix, the bases are in the inside and are connected to their complementary base (adenine to thymine and cytosine to guanine) by hydrogen bonds. The two strands of the DNA double helix run in opposite directions, one in the 5' to 3' direction, know as the Watson strand, and the other in the 3' to 5' direction, known as the Crick strand. The strands relate to each other in a antiparallel way.

Figure 1.3: Double helix model (from [12]).

1.2 Chargaff's rules

Erwin Chargaff was a biochemist that through careful experimentation discovered a few rules that helped lead to the discovery of the double helix structure of DNA. His four rules are [11]:

- First parity rule: For duplex DNA the percentage of adenine is equal to the percentage of thymine, and cytosine and guanine share the same characteristic.

$$\%A = \%T, \%C = \%G \tag{1.1}$$

- The cluster rule: Individual bases are clustered more than it could be expected on a random basis.
- Second parity rule: For individual strands of DNA, to close approximation, the first parity rule also applies.

$$\%A \cong \%T, \%C \cong \%G \tag{1.2}$$

- The GC rule: the ratio of $C+G$ to the total bases $(A+C+G+T)$ tends to be constant in a particular species, but varies between species.

The cluster rule, first and second parity rules are species-invariant, while the GC rule is species-variant.

1.3 Oligonucleotides

Oligonucleotides are a sequence of nucleotides. Since there are 4 different nucleotides in DNA, the number of different possible oligonucleotides with N nucleotides are $n = 4^N$. Oligonucleotides of order two and three and generally referred to as dinucleotides or doublets and trinucleotides or triplets respectively.

1.4 Oligonucleotide conjugates

There are three different oligonuleotide conjugates (reverse, complement and inverted).

A reverse conjugate is the reverse of the oligonucleotide. An oligonucleotde with 5' to 3' direction has a reverse conjugate which is the same sequence with 3' to 5' direction.

Complement conjugates explore the base pair complementarity, adenine with thymine and cytosine with guanine. The complement conjugate of an oligonultotide is one that for each original nucleotide it contains it's complementary pair.

The inverse conjugate is the reverse of a complemented oligonucleotide.

To illustrate each of the three types of oligonucleotide conjugates, the following oligonucleotide of order $N = 5$ is used as an example:

$$AAGTC \tag{1.3}$$

The oligonucleotide conjugates are:

- Reverse conjugate: $CTGAA$.
- Complement conjugate: $TTCAG$.
- Inverse (also known as inverted complement): $GACTT$.

These different conjugates are the base of our symmetry analysis. Reverse, complement and inverse symmetry depends on the quantities of these conjugates respectively. A genome with a perfect symmetry is one where every oligonucleotide is present in the same quantity as it's conjugate.

1.5 State of the art

Several studies and research have been conducted in the field of DNA symmetry. The first Chargaff's parity rule (that states that for duplex DNA the percentage of adenine is equal to the percentage of thymine, and cytosine and guanine share the same characteristic) is well documented [11], as well as the second parity rule (that for individual strands of DNA, to close approximation, the first parity rule also applies) [5].

These studies focused on quantitative results for single nucleotide ocurrence.

Trying to generalize to higher orders (i.e. comparing the quantities of oligonucleotides instead of nucleotides) several studies have been conducted as well. In [4, 20] oligonucleotides of order up to $N = 10$ were compared, while others [2, 3, 15] focused on dinucleotide and trinucleotide frequencies. All of these studies concluded that there is up to some order inverted symmetry. Kong *et al* [15] suggest inverse duplication (figure 1.4) as an hypothesis to the genome growth and evolution, resulting on the found symmetry.

Reverse and complement symmetry also has been studied, but results have not shown reverse or complement symmetry at any order [15].

In order to quantify symmetry levels, several measurements have been used, from relative to global measurements. Relative measures focus on measuring symmetry between specific nucleotides or oligonucleotides and their conjugate, while global measures focus on a global measurement for all oligonucleotides and their conjugates of a given DNA sequence. Relative

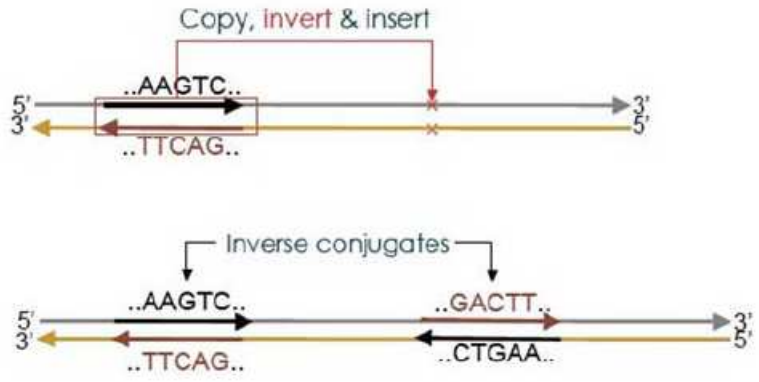

Figure 1.4: Inverted segmental duplication for DNA genomes, from [15].

measures previously used to measure DNA symmetry are frequency skews [4] and relative abundance [22]. Global measures used include L1 distance [15], KullBack-Leibler divergence [4] and correlation coefficients [2].

In order to graphically represent DNA symmetry, plotting the frequency of all possible oligonucleotides and the frequency of their conjugates at a given order has been done (figure 1.5). This method allows us to clearly see if symmetry is present. The closer the points are to the line $y{=}x$, the more symmetric the sequence analysed is.

Another method used to graphically represent DNA symmetry is to plot all the different possible oligonucleotides and their respective frequency (figure 1.6). This method may requires us to mentally calculate the oligonucleotide conjugates in order to compare their frequency if the the x axis is not ordered by conjugates as is the case of (figure 1.6). At high orders this can be almost impossible.

All of these studies concluded that as a general rule, the bigger the sequence analysed, the higher the symmetry level is. For large DNA genomes, such as the human genome, even at the higher oligonucleotide orders analysed ($N{=}10$), inverse symmetry is present [4]. This led us to try to find at witch oligonucleotide order does inverse symmetry stop. Beside the relative low order of the analysis done, some of these studies also suffer from limited set of sequences or lack of species and measurement diversity. This led us to implement a extensive tool capable of:

- Analyse DNA sequences at high oligonucleotide orders to determinate at what order does inverted symmetry stop for large DNA genomes.

- Analyse all three types of symmetry (inverse, reverse and complement) in order to confirm the absence of reverse and complement symmetry.

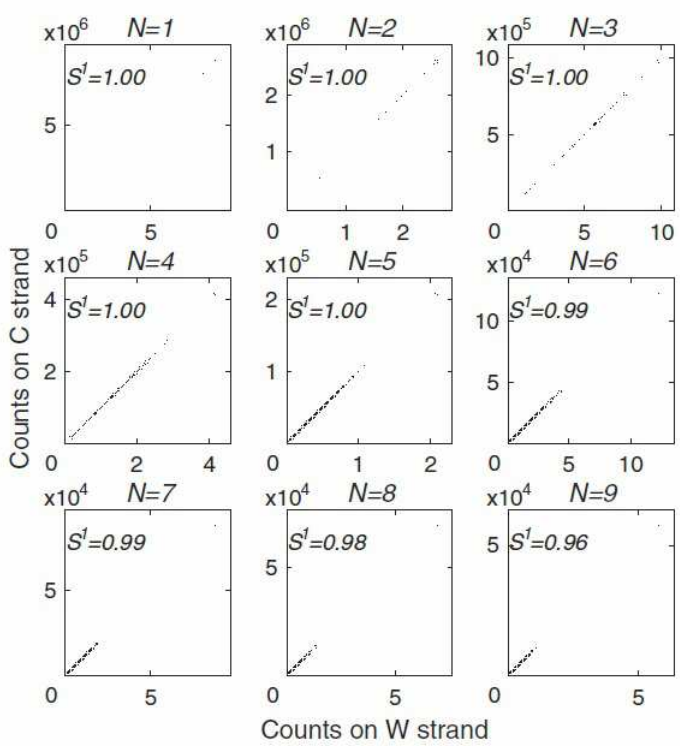

Figure 1.5: Counts on direct strand vs counts on reverse complementary strand of human cromossome 22 of oligonucleotide lengths of $N = 1$ to 9 (from [4]).

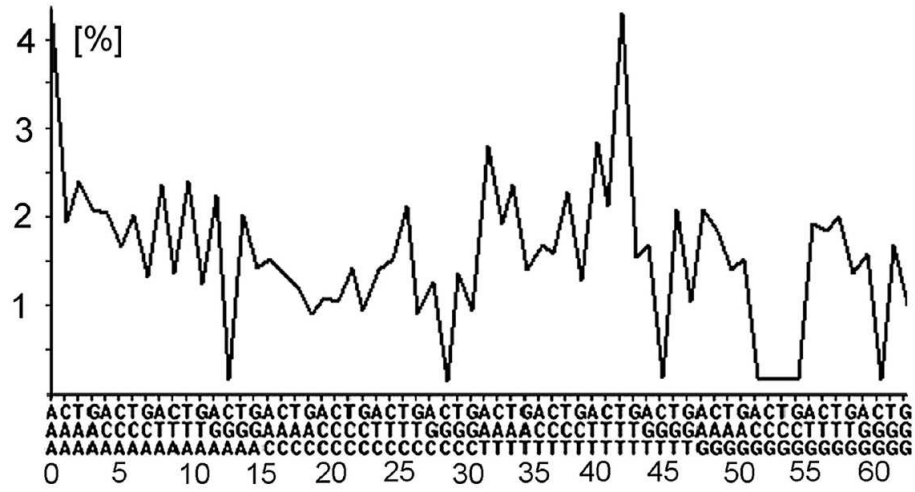

Figure 1.6: Trinucleotide frequencies in percentage (from [3]).

- Quantify symmetry using diverse measures and find which ones are better to classify DNA symmetry.

Furthermore a large variety of DNA sequences from diferent species of different families will be analysed such as archaea, bacteria, protozoa, fungi, insects, mammals, nematodes, fish and plants, in order to try to find a relation between them.

Chapter 2

Information retrieval

Processing large quantities of information can be very onerous, both in time and computer memory. In this chapter the process of extracting information will be described, including file processing, and the complexity of data structures will be analysed.

Big O notation will be used to describe the complexity, more specificity O() for the worst-case scenario and $\tilde{O}()$ for the average case scenario.

The main challenges, programming language wise, are the ability to process large data files, its portability, and the possibility of future implementation of rich graphic interface. The programming language we have chosen was java [17] because of the author's experience with it and the fact that it can tackle all of the above requirements. C and C++ where also considered, but the author's lack of experience with C++, the weaker graphical interfaces and its portability narrowed down the pick to java.

2.1 FASTA specification

FASTA is a DNA and protein sequence alignment software package developed by David J. Lipman and William R. Pearson in 1985. Their legacy is the FASTA format which is now ubiquitous in bioinformatics.

FASTA format is a text based format that can represent both nucleotide sequences and peptide sequences, using a single letter code to represent base pairs or amino acids.

A sequence begins with a single line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than ">" symbol at the beginning. It is recommended that all lines of text are shorter than 80 characters in length [14]. An example sequence in FASTA format is shown in figure 2.1.

FASTA specifications do not allow blank lines in the middle of the file input. A multiple sequence is allowed, obtained simply by concatenating several single sequence FASTA files.

```
>gi|221070640|ref|NW_002238186.1|Vvi3_WGA32_1 Vitis vinifera chromosome 3 genomic contig
TAATTGAGTACTTAAAATAAAAACTTTTCATCCATCCTCTTTCTTTCTTTTTCTTTTTCTTATTTTAATA
AATTTTCAATTAATTCAAATCATTAAAATAAATCCTTAATAAACAAATTTGTAACATTTCATATACCTTA
.
.
.
>gi|221070635|ref|NW_002238181.1|Vvi3_WGA27_1 Vitis vinifera chromosome 3 genomic contig
AAATATGTGTATGTGACATCATTTAATTAGCTGTGCTTTTTCGATAAGTATTGAAAAAATATATTTTTAT
ATGGTAGTTGGATTTGAAAATTTTGGATAAACCAACTTCTTACCTAATAGTAAATAAACAAATAAATCAA
ATAAGGCGTTTACTTATAACCTAGTCTATCTTCAATTTCTTATATAAGCATTTATGAATGCATATTTGTT
.
.
.
```

Figure 2.1: Example of a file in FASTA format (Vitis vinifera chromosome 3).

This sticks to the format as the first line of a file starts with ">", forcing all subsequent sequences to start with a ">" in order to be taken as different ones. The example shown in figure 2.1 is in fact a multiple sequence FASTA format file.

The files contain the nucleotides, A(adenine), C(cytosine), G(guanine), T(thymine), as well as N. N represents an unknown nucleic nucleotide, and we parse them out as well as the header lines for our information retrieval.

Header lines normally contain accession versions (e.g. AA01.1) but a bar-separated NCBI sequence identifier (e.g. gi|129543) is also acceptable. These NCBI sequence identifiers have a specific syntax [9] and generally contain information about database location, patent information or dates. For example, for a patent type, the format is (pat|country|patent|sequence; example:pat|US|RE33188|1).

There is no file extension standard in FASTA specifications but the most commonly used are:

- *.fa - generic fasta file.
- *.fna - contains coding and non coding regions of a genome.
- *.fnn - contains coding regions of a genome.
- *.faa - contains amino acids.
- *.frn - contains non coding RNA regions of a genome.

For our purposes we are using fa, fna and fnn files.

2.2 Mapping values

In order to process large DNA files at high oligonucleotide orders, we need to identify each oligonucleotide as fast as possible, as well as use the minimum physical memory possible to store information about them (i.e the number of ocurrences).

Each character in java language requires 16 bits, so for each oligonucleotide of order $N = 20$ it would take 320 bits of memory to save it as a series of characters.

Mapping each possible oligonucleotide to an unique numerical number, using a long 64 bit data type, we can identify oligonucleotides up to orders of $N = 31$, since a long data type in java holds values up to $2^{63} - 1$. Mapping them to an integer 32 bit type will allow up to orders of $N = 15$. This solution also proves to be much faster for comparing oligonucleotides.

Identifying each possible oligonucleotide is done by the following formula:

$$\sum_{i=0}^{N-1} F_i(4^{N-i-1}) \tag{2.1}$$

F_i is defined by:

- 0 when the character in position i is A.
- 1 when the character in position i is C.
- 2 when the character in position i is G.
- 3 when the character in position i is T.

Given the following sequence for better understanding:

$$ACGAC \tag{2.2}$$

Processing this sequence for order $N = 2$ would generate the following numbers for each of the dinucleotides present in the sequence:

- AC -> $F_A(4^{2-0-1}) + F_C(4^{2-1-1}) = 0(4) + 1(1) = 1$.
- CG -> $F_C(4^{2-0-1}) + F_G(4^{2-1-1}) = 1(4) + 2(1) = 6$.
- GA -> $F_G(4^{2-0-1}) + F_A(4^{2-1-1}) = 2(4) + 0(1) = 8$.
- AC -> $F_A(4^{2-0-1}) + F_C(4^{2-1-1}) = 0(4) + 1(1) = 1$.

2.3 Data structures

The storage and organization of our information in memory must be done in the most efficient way possible as processing DNA files is very demanding in terms of time and memory. Different data structures are suited to different type of applications. We will now explore some data structures (simple arrays, dynamic arrays, linked lists, binary trees and hash tables) and try to find the most efficient ones for our specific task. We will focus on their efficiency and complexity when doing the operations of insertion and search, and we will not deal with deletion because they are not needed for our work.

2.3.1 Simple arrays

A simple array of an object in a computer memory is typically addressed by integers from 0 to $N-1$, where N is normally an integer. In most programming languages, an object occupies a contiguous set of locations in the computer memory. It uses a pointer, that is the address of the first memory location of the object, and we can address other memory locations within the object by adding an offset to the pointer (index). For our task, the objects are 32 bit integers. These integer objects are the number of counts of each possible oligonucleotide, while the indexes themselves represent the different possible oligonucleotides (oligonucleotide identifiers). This allows us to only store in memory the number of counts and nothing else making the single array the most efficient possible solution whenever all the possible oligonucleotides have occurred at least once, because we only store the 32 bits corresponding to the oligonucleotide counts and not the 64 bit of the oligonucleotide identifiers. As we can see in table 2.1 this solution is not feasible and is a problem for high oligonucleotide orders because of the amount of memory required.

Given the sample DNA sequence (2.2), analysing the sequence for oligonucleotide order $N = 2$, the state of memory is shown in table 2.2.

Complexity wise, using simple array will also be the most efficient, as both insertion and searching operations behave in a O(1) manner, since neither of the operations are affected by the size of the structure.

2.3.2 Dynamic arrays

Dynamic arrays differ from simple arrays because the array size is not static, having the ability to change size after the array's declaration. It allows elements to be added or removed. Most modern programming languages include them in their standard libraries. Dynamic arrays are constructed with a fixed size array that consists of two parts. The first, consists of the used portion of the dynamic array and the second part of the free space. As the size of the logical

N	Size
1	16B
2	64B
3	256B
4	1KB
5	4KB
6	16KB
7	64KB
8	256KB
9	1MB
10	4MB
11	16MB
12	64MB
13	256MB
14	1GB
15	4GB
16	16GB
17	64GB
18	256GB
19	1TB
20	4TB

Table 2.1: Oligonucleotide N order and memory required for single array.

index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
content	0	2	0	0	0	0	1	0	1	0	0	0	0	0	0	0

Table 2.2: State of the memory used using simple array with sample sequence $ACGAC$ for oligonucleotide order $N = 2$.

size (used portion) grows, the capacity itself has to grow in order to accommodate more entries. This resize of the array occurs whenever the space is entirely consumed and the array cannot hold more entries. Resizing the array is a very expensive task, typically involving copying the entire contents of the array, so choosing the initial capacity is of extremely importance.

The dynamic array can solve the problem of the simple array in terms of memory used

because the oligonucleotides that never occurred in our DNA sequence do not need to be allocated in memory. In figure 2.2 we can see how the insertion of the sample sequence (2.2) would occur.

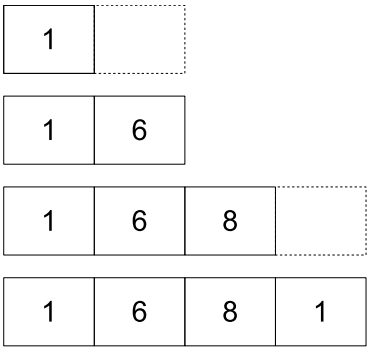

Figure 2.2: Unsorted insertion of the sample sequence $ACGAC$ for oligonucleotide order $N = 2$ in dynamic array.

Using an unsorted dynamic array, as in 2.2, would only require 64 bit per each entry, but instead of allocating the number of occurrences of the olygonucleotides we would be allocating the olygonucleotide identifiers, requiring us to allocate the same olygonucleotide identifier several times. Depending on the size of the DNA sequence this solution can be several times worse in terms of memory required, because the same identifier would be stored several times (the number of time it appears in the DNA sequence).

Allocating both oligonucleotide identity and number of occurrences would take 96 bits per entry but would save overall much more memory (figure 2.3), whenever the the same identifier is present several times in the DNA sequence, because it wouldn't have to be stored more than once.

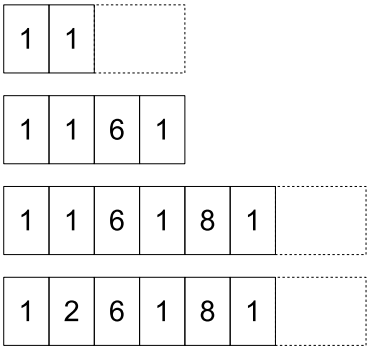

Figure 2.3: Sorted insertion of the sample sequence $ACGAC$ for oligonucleotide order $N = 2$ in dynamic array.

Inserting new elements at the end of the dynamic array can be done in constant time (O(1)), but we need to keep the dynamic array sorted to save memory, so an insertion requires a previous search. Implementing a binary search would require in the worst case O(log(M)) time (M representing the number of different oligonucleotides present in the DNA sequence). Inserting an element after the search is done takes O(M) time because in order to keep the array sorted, every single entry after the right position is found has to be changed. So the operations complexity is:

- Insert: O(M)+O(log(M)).
- Search: O(log(M)).

In spite of the memory improvements over the simple arrays (when not all oligonucleotides are present in the DNA sequence, because oligonucleotides not present in the sequence do not have to be stored), dynamic arrays fail at delivering reasonable insert and search time, making it a solution not adequate for this work.

2.3.3 Linked lists

A linked list is a data structure in which the objects are arranged in linear order. Instead of the linear order being determined by the array indices, like in a simple array, the order in a linked list is determined by a pointer in each object.

In doubly linked lists, an object has two pointer attributes: one to the next and another to the previous object. The object may contain any data we require. Given an element in the list, the next pointer points to its successor in the linked list, and the previous pointer points to its predecessor. If the previous pointer is NULL, then that element does not have an predecessor and is therefore the head (first element). Similarly if the next pointer of an element is NULL it is the tail (last element of the list).

Processing the sample sequence (2.2) for oligonucleotide order $N = 2$, using an ordered by olygonucleotide double linked list containing both olygonucleotide identity and occurrence count, would result in the structure shown in figure 2.4.

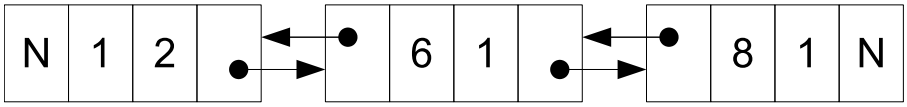

Figure 2.4: Double linked list after inserting sample sequence $ACGAC$ for oligonucleotide order *N=2*.

In comparison with dynamic arrays, linked lists require more memory since we are allocating the next and previous pointer for each element. But this allows the insertion of an element to be much faster as we no longer have to move every single element after finding the position to do the insertion. We simply change the next pointer of the previous element and the previous pointer of the next element pointing them to the new element. One of the problems of implementing linked lists is that we can no longer use binary searches. We cannot find the middle of the list in constant time, resulting in O(M) searching times. The complexity of the operations is:

- Insert: O(M).
- Search: O(M).

Linked arrays greatly improve insertion times, but we simply cannot afford to do linear searches.

2.3.4 Binary trees

A binary tree is a linked data structure where each node is an object. In addition to a key (primary tribute that defines the weight on the node, in our case the oligonucleotide identifier) and satellite data (other information stored, in out case the oligonucleotide counts), each node contains attributes left, right, and parent that point to the nodes corresponding to its left child, its right child, and its parent, respectively. Whenever a child or the parent is missing, the corresponding attribute contains the value NULL. The only node that has no parent is the root node.

A binary search tree is a binary tree that respects the following properties:

- Every key in the node's left subtree is smaller than the node's key.
- Every key in the node's right subtree is bigger than the node's key.

Basic operations on a binary search tree take time proportional to the height of the tree. In complete binary trees (when the height distance of the root node is the same to all end nodes) basic operations are done in O(log(N)) time. However, if the tree is a linear chain of nodes (figure 2.5), operations are done in a linear way (O(M)).

A balanced binary search tree is height balanced. For every node, the heights of the left and right subtree cannot differ by more than 1. The constant need of balancing the tree is a very expensive task as inserting a new node can change all the whole structure. Figure 2.6 shows the state of the balanced tree after processing the sample sequence (2.2).

The complexity of the operations is:

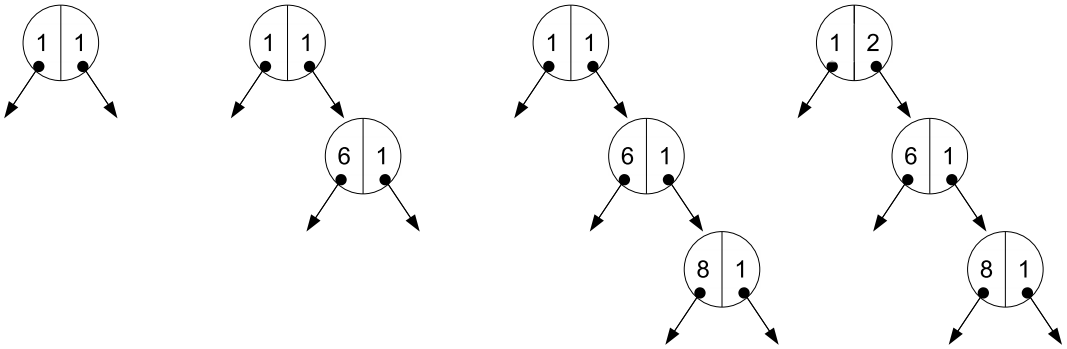

Figure 2.5: State of an unbalanced binary tree after inserting the sample sequence $ACGAC$ for oligonucleotide order $N = 2$.

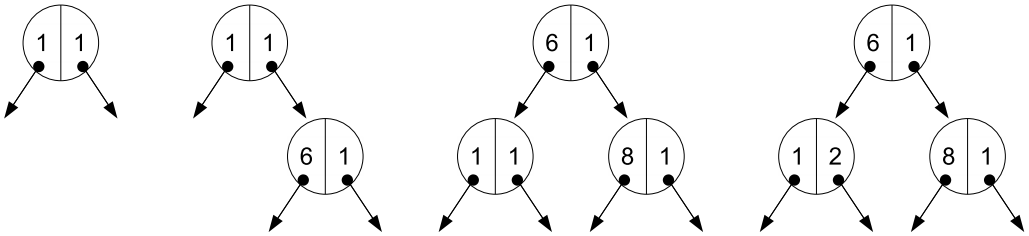

Figure 2.6: State of a balanced binary tree after inserting the sample sequence $ACGAC$ for oligonucleotide order $N = 2$.

- Insert: O(M).
- Search: O(log(M)).

In spite of the constant balancing need, binary trees improve on search time comparing to linked lists.

2.3.5 Hash tables

A hash table generalizes the simple notion of an ordinary array. As we have seen with simple array, directly addressing makes effective use of our ability to search an arbitrary position in an array in O(1) time. The downside is that when the keys to be store are actually a small number compared to the total possible keys and we cannot afford to allocate the whole universe of keys, because of the amount of memory required.

Hash tables are an efficient alternative to directly addressing an array, typically using an array of size proportional to the number of keys stored. Instead of using the key as an array index directly, the array index is computed based on the key. With a simple array, an element with key k is stored in slot k. With hashing, this element is stored in slot f_k, using a hash function f to compute the slot from the key k. Figure 2.7 shows a simple example when f_k equals the key k and the sequence processed is (2.2) and oligonucleotide order $N = 2$ without collisions.

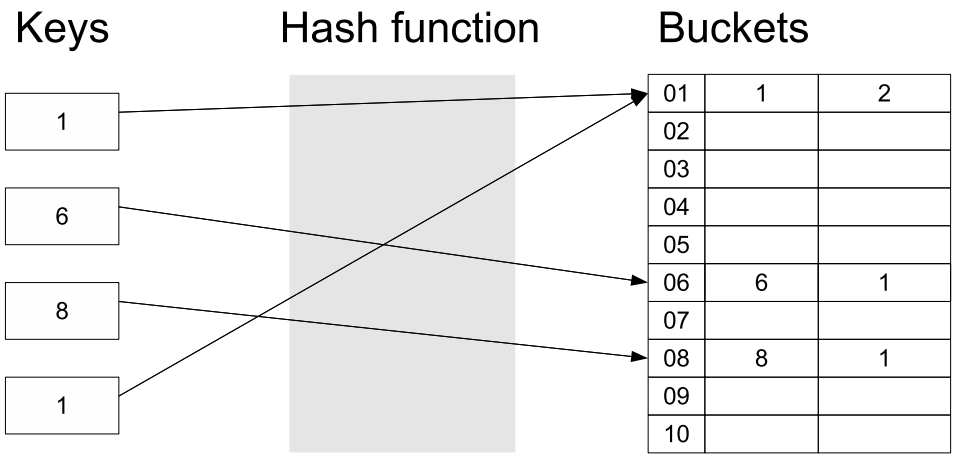

Figure 2.7: Hash table without collisions, insertion of the sample sequence $ACGAC$ for oligonucleotide order $N = 2$.

The function F_k is calculated based on the key but not only, the actual size of the hash table, creating the possibility of collisions. This happens when more than one key hashes into the same bucket. One way to solve collision problems is by chaining (figure 2.8).

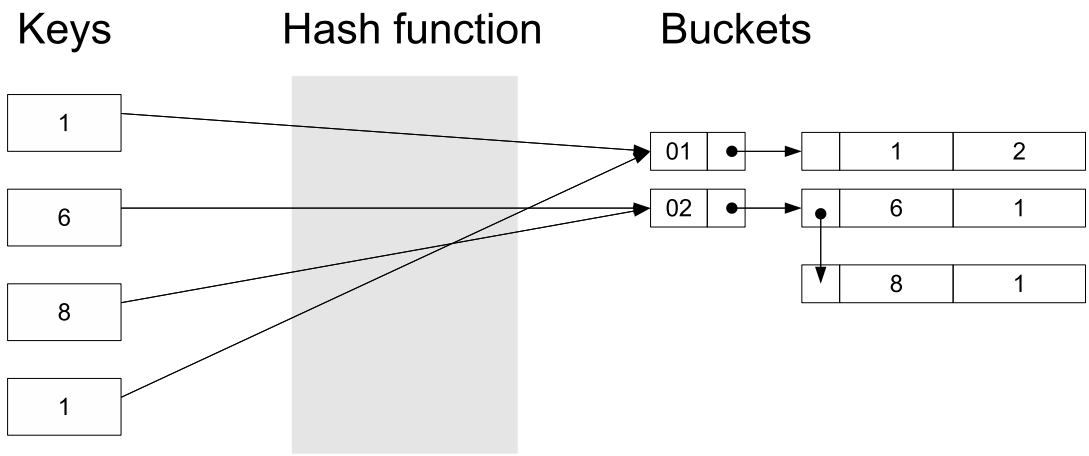

Figure 2.8: Hash table with collisions, insertion of the sample sequence $ACGAC$ for oligonucleotide order $N = 2$.

Chaining uses a linked list to place all the elements that hash into the same bucket. If the bucket does not contain any elements, it contains null. The number of collisions are directly dependant on the size of the hash table, making the load factor (the portion of the buckets in the structure that will be used by one or more entries) very important on hash table performance. Each time the hash table goes over the load factor threshold it grows in size. Using a high load factor will result in less memory allocation but will increase the chances of collisions. As the load factor approaches zero, the probability of collision occurring diminishes but the memory needed increases as the portion of unused buckets is bigger. Initial capacity is also important, as rehashing the table is time consuming.

The complexity of the operations is[7]:

- Insert: $\tilde{\mathrm{O}}(1)$.
- Search: $\tilde{\mathrm{O}}(1)$.

2.4 Solution adopted

After studying the various data structures we have concluded:

- Whenever the whole universe of keys is present and the memory is sufficient, the simple array solution is the most efficient both in terms of memory and complexity.
- We cannot afford to use simple arrays for high oligonucleotide orders because of the size of memory needed (table 2.1).

These constraints of the simgle array led us to implement another type of data structure in order to process DNA files using high oligonucleotide orders.

N	Insertion	Search
Simple array	O(1)	O(1)
Dynamic array	O(M)+O(log(M))	O(log(M))
Linked list	O(M)	O(M)
Balanced binary tree	O(M)	O(log(M))
Hash table	$\tilde{O}(1)$	$\tilde{O}(1)$

Table 2.3: Insertion and search complexities for simple array, dynamic array, linked list, balanced binary tree and hash table.

Beside simple arrays, hash tables are the most efficient data structure for our task (table 2.3). It still takes longer than simple array to do a search and an insertion, because of the rehashing and the collisions. But under reasonable assumptions it still averages constant search and insertion times.

Having chosen both simple arrays and hash table to implement, one question rises. When should we use simple array and when should we use hash table to process our DNA files?

Table 2.4 shows the time efficiency for both simple array and hash table implementations for oligonucleotide N orders of up to 15, processing Ecoli bacteria DNA files. Holes represent the number of keys in the universe that never occurred. Processing file time corresponds to the time it took to count all oligonucleotide occurrences in our DNA files. Retrieving values are the time it took, after all the insertions were done, to search all the oligonucleotide identifiers present and calculate symmetry and similarity values. The times represent the runtime of the university server (*Sapiens*) , consisting of four quad core intel xeon e7320 processors running at 2.13GHz of speed, with 256GB of shared RAM memory.

As expected, while all the keys of the universe were present, using a simple array is faster ,or as fast approximately both on processing and retrieving times. But with the growth of the N order, holes start to appear and its numbers keep increasing for higher orders. Processing time continued to be faster using a simple array, but retrieving values is much faster using a hash table. For $N = 15$ the hash table performed approximately 130 times faster on retrieving the values than the simple array, while being only less than twice slower processing the file. This happens because the simple array has to iterate over all the universe of keys, even if they are not present in the sequence processed, while the hash table only has to iterate over the present keys.

N order	Holes	Simple array processing file(ms)	Hash table processing file(ms)	Simple array retrieving values(ms)	Hash table retrieving values(ms)
1	0	600	977	6	6
2	0	590	977	6	6
3	0	672	1154	7	7
4	0	756	1396	9	9
5	0	884	1599	21	21
6	0	950	2021	70	67
7	1	1039	2661	145	151
8	176	1136	2973	302	411
9	5617	1378	3466	940	985
10	150468	1764	4219	2969	2846
11	1997469	2039	5014	11835	8317
12	13298293	2240	5426	47566	15643
13	62938541	2497	5444	185741	19579
14	264003264	2817	5513	732498	22480
15	1069224203	3098	5806	3115171	24091

Table 2.4: *Ecoli bacteria*, time efficiency comparison between simple array and hash table.

Calculating *a priori* the number of holes, in order to choose what data structure to use is impossible. The only way would be to iterate over the DNA file previously, and that would be very time demanding. For that reason we calculate the total number of oligonucleotides occurrences using the file size and use it as an upper bound to choose what data structure to use. Whenever the total number of oligonucleotide present in the sequence is lower than the universe of keys, holes will necessarily occur, making us choose the use of a hash table instead of a simple array.

Chapter 3

Data compression

Processing large genome files (several GBs of size) requires a large amount of time, several hours or even days, depending on the size of the genome, the oligonucleotide order and the speed and memory of the computer used.

Comparing the time it takes to iterate over the files, identify the oligonucleotides and store the number of occurrences in memory (processing file time) with the time to retrieve each unique oligonucleotide present and calculate symmetry and similarity measures (retrieving values time), we can clearly conclude for oligonucleotide orders below 15, that the processing time has a bigger impact on the overall time of evaluating symmetry, than the retrieving time (table 3.1). One of the main causes for the high retrieving time values is that in order to identify an oligonucleotide of order N, the power function has to be invoked N times. Processing one GB DNA sequence with oligonucleotide order of 10 would results in invoking the power function 10737418231 times.

When looking at the equation to identify each oligonucleotide (2.1) we can see that the power function is always base 4, and when processing a DNA sequence for oligonucleotide order N, to identify each oligonucleotide the exponent of the power function varies from 0 to $N-1$. This allows us to calculate and store at the start of the file processing the values of the power function with base 4 and with N from 0 to $N-1$. Doing this, we only have to execute the power function N times for the whole analysis, and point to that value whenever needed. Table 3.2 shows the time required to identify 2^{24} oligonucleotides of order 30, using the stardard java library power function, referencing the values previously calculated, as well as a specific power function of base 4 that we implemented using shifts.

As observed in table 3.2, the fastest method is to calculate previously the power functions and pointing to the values whenever needed. The base 4 function using shifts also performed well, but not as fast. The java library power function performed poorly in comparison.

To minimize processing times further, we decided to add a save and load functionality to

N order	Processing file(ms)	Retrieving values(ms)
1	162423	6
2	175781	5
3	223014	5
4	279880	8
5	331428	20
6	384179	64
7	451659	219
8	499910	263
9	774981	759
10	1014128	2672
11	1143294	11342
12	1222768	46437
13	1389894	193157
14	1624510	793706
15	1768216	3274338

Table 3.1: Time efficiency for simple array, processing *Homo sapiens* complete genome.

Java library power function (ms)	229976
Specific power function of base 4 using shifts (ms)	2038
Referencing the power function previously calculated (ms)	1290

Table 3.2: Time efficiency for different power function methods, calculated by identifying 2^{24} oligonucleotides of order 30.

our tool, in order to save the information of our data structure, removing the processing time of future analyses of the same genome. This requires to save oligonucleotide counts whenever we use simple arrays or saving both oligonucleotide counts and oligonucleotide identifiers when we are using hash tables. One problem is, the sheer size of the files created in the process (table 3.3), leading us to encode the results, using fewer bits than the original representation would use.

N order	Holes	Simple array uncompressed size	Hash table uncompressed size
1	0	16B	48B
2	0	64B	192B
3	0	256B	768B
4	0	1KB	3KB
5	0	4KB	12KB
6	0	16KB	48KB
7	0	64KB	192KB
8	0	256KB	768KB
9	0	1MB	3MB
10	0	4MB	12MB
11	979	16MB	47.99MB
12	167126	64MB	190,09MB
13	4801682	256MB	713,05MB
14	65721695	1GB	2,32GB
15	527172585	4GB	6,25GB

Table 3.3: *Homo sapiens* complete genome, uncompressed size of the oligonuleotide information store on simple array and hash tables.

3.1 Compression methods

Data compression is the process of encoding information using fewer bits than it would originally require.

When we are speaking of compression methods, we are actually referring to two different processes. The compression process and reconstruction, the first takes an input x and represents it as y, requiring less bits, the second process takes the representation y and reconstructs it as z (figure 3.1).

Based on the reconstruction requirements we can classify compression methods in two types, lossless compression when the reconstructed information z is equal to original information x and lossy compression when z and x are different. Lossy compression methods achieve better compression ratios at the price of losing some information. When we are recovering the information compressed, the result is not exactly the original data stream. Such methods make sense especially when compressing images, video, or audio, if the loss of data is small, we

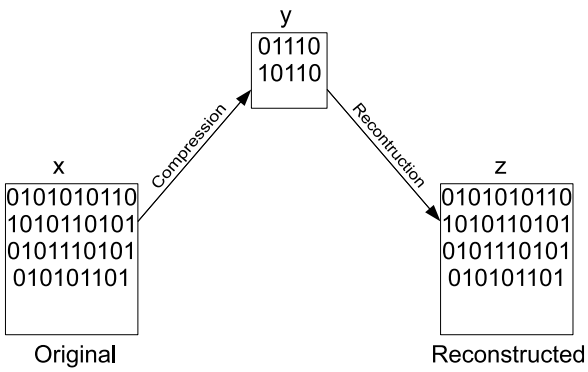

Figure 3.1: Schematic representation of compression and reconstruction.

may not be able to tell the difference. In contrast, text files or files containing computer programs, may become worthless if even one bit gets modified. Such files should be compressed only with a lossless compression method.

For our task we need to recover the exact same values that we compressed, in order to measure symmetry, so we will focus on exploring only lossless compression methods.

3.2 Predictors

Reconstruction requirements may force us to decide whether the compression needs to be lossy or lossless, but the exact compression scheme that will be used depends on a number of different factors. The characteristics of the data that needs to be compressed are one of the most important factors. A compression technique that works well for the compression of video may not work well for compressing sound. Each application presents a different set of challenges.

The approach that works best for a particular application will depend to a large extent on the redundancies inherent in the data.

The development of compression algorithms can be divided into two phases. The first phase is usually referred to as modelling or predicting. In this phase we try to extract information about any redundancy that exist in the data and describe the redundancy in the form of a model. The residual is often referred to as the difference between the data and the model.

During the compression phase the prediction is done before the coding (figure 3.2). When recovering the information, the decoding algorithm is run before the predicting is taken into account (figure 3.3).

We need to compress the oligonucleotide counts when using simple arrays and both oligonucleotide counts and oligonucleotide identifiers when using hash tables. As an example, we

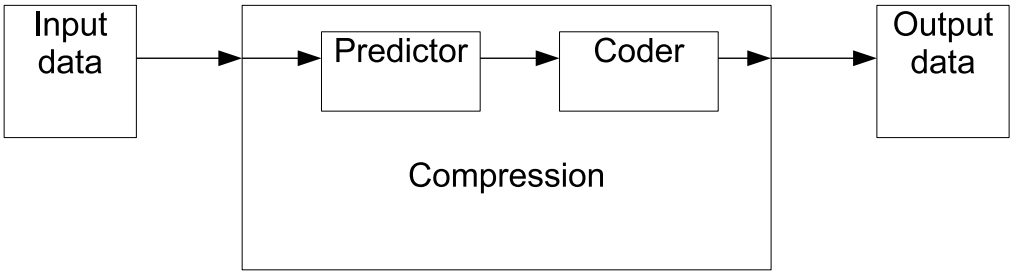

Figure 3.2: Block diagram of the compression phase with a predictor.

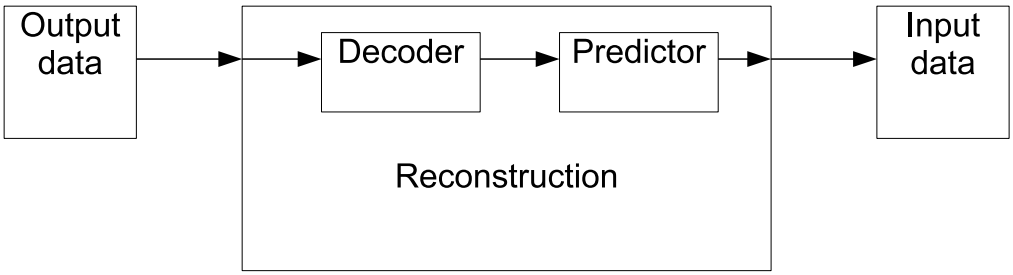

Figure 3.3: Block diagram of the reconstruction phase with a predictor.

describe the processing of a DNA sequence for oligonucleotide order $N = 2$, with all oligonucleotides present, except TA would result in the following identifiers:

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15.

When looking at these identifiers we can clearly see a relation between then. Using the predictor

$$\text{residual} = \text{actual} - \text{previous} + 1 \tag{3.1}$$

would give us the following values:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0.

These values require less bits than the original identifiers. For oligonucleotides of higher orders, residuals are not so small because of the high number of holes but, even so, the number of bits required to represent these residuals is lower than the original values.

Oligonucleotide counts behave in the opposite manner. When the number of holes is high, the number of those counts is much smaller than when there are no holes. Higher N order oligonucleotides result in smaller counts, per oligonucleotide. Let us assume the oligonucleotide counts of the previous DNA identifiers are:

20, 7, 10, 16, 22, 8, 5, 4, 28, 31, 16, 8, 1, 16, 18.

Predicting the oligonucleotide counts is not as easy, but if we use the predictor

$$\text{residual} = \text{round}(\text{mean} - \text{actual}) \tag{3.2}$$

results in the following values:

mean = 14; values = -6, 7, 4, -2, -8, 6, 9, 10, -14, -17, -2, 6, 13, -2, -4.

Other statistic models such as median could possibly result in better predictors, but we can calculate the mean without having to iterate over the structure, using the total number of oligonucleotides, possible oligonucleotide identifiers and number of holes if needed. Calculating other statistics would be less time efficient, requiring us to iterate the structure. Improving time efficiency led us to data compression and not the other way around.

The formula for calculating the mean of the oligonucleotide counts when using simple array data structure is different from that used with the hashtable, because simple arrays store all the possible oligonucleotide identifiers counts, while the hashtable only stores the oligonucleotide counts that were present in the processed DNA sequence. When using simple arrays, the mean of the oligonucleotide counts can be calculated by:

$$\text{mean} = \frac{\text{total oligonucleotide count}}{\text{possible oligonucleotide identifiers}} \tag{3.3}$$

When using hash tables, the mean of the oligonucleotide counts can be calculated by:

$$\text{mean} = \frac{\text{total oligonucleotide count}}{\text{possible oligonucleotide identifiers} - \text{number of holes}} \tag{3.4}$$

3.3 Lossless data compression

3.3.1 Max value data compression

Compression methods can be classified as fixed length coding or variable length coding. Fixed length coding uses the same number of bits to represent all the different possible symbols. Variable length coding uses a different number of bits to represent each symbol, in order to take advantage of using less bits to represent the symbols that have the highest probability to occur.

Max value data compression is a very simple method of fixed length coding that codes all the values with the number of bits required to compress the maximum value present in the dataset.

When compressing the oligonucleotide counts, if the maximum difference from the mean of the values requires 16 bits to represent, this would give us a 0.5 compression ratio (3.5) because we are using 32 bit integer variables to allocate the oligonucleotide counts. In this

work we use the following definition of compression ratio:

$$\frac{\text{compressed size}}{\text{uncompressed size}} \tag{3.5}$$

3.3.2 Arithmetic data compression

Arithmetic coding [10] is a variable-length data compression technique that instead of separating the input message into its component symbols and replace each symbol with a code word, it encodes the entire message into a single number.

Arithmetic coding uses a table of probabilities. The single number to be encoded will be a value inside the probability range [0,1). All probabilities fall into the range [0,1) while their sum equals 1 in every case. The [0,1) interval is partitioned according to the probability distribution of the symbols and then, after iterating this step for each symbol in the message, a value inside the final interval is chosen for representing the message. To decode the message, the encoder is applied backwards.

There are two ways of applying the table of probabilities, which is also known as finite-context models: statically and dynamically. Static methods requires previous knowledge and storage of the probability information. With dynamic methods, the statistical information is adapted as the coding processes.

Arithmetic coding is especially useful when dealing with sources with small alphabets, and alphabets with highly unequal probabilities. Because of the small DNA alphabet, consisting of only 4 symbols (A, C, G, T), arithmetic coding has been used [19]. In our case, we are compressing the oligonucleotide identifiers and not the actual sequence. Transforming these oligonucleotide identifiers back to oligonucleotide sequences would defeat the purpose of minimizing processing times. It would also require to save two separate files when using hash tables, one for the oligonucleotide identifiers and one for the oligonucleotide counts.

3.3.3 Dictionary data compression

In many applications, the input data consists of recurring patterns. Dictionary compression methods are based on constructing a structure containing the patterns that appear most frequently. Dictionary's can be static or dynamic (adaptive). Static dictionary's are permanent, sometimes permitting the addiction but not deletions from the structure, while dynamic dictionary's store patterns previously encountered in the input stream, allowing both addictions and removals. A static dictionary is only appropriate when prior knowledge about the source is available. Dynamic dictionaries start with an empty dictionary or very small (default) dictionary, and add patterns to it as they are found in the input stream. It can delete old words to maintain the size of the dictionary under control. This method consists on a loop

reading the input stream and then search the dictionary to find if the pattern is found. If it is found, then it writes the index of it on the output stream. Otherwise, the pattern is added to the dictionary. Examples of dictionary compressors are LZ77 and LZ78 [25]. Specific to our task this method would require us to use more memory because of the dictionary construction necessity and would be time consuming, looking up the values in the dictionary. Additionally when encoding small numbers, the index itself could possibly require more bits than the original number. Some dictionary compressors additionally use other codes in order to use lower length code words for values that appear with higher probability. This would result in even more time.

3.3.4 Golomb coding

Solomon Wolf Golomb is an American mathematician and engineer, known for inventing polyominoes, the inspiration for Tetris, as well as the Golomb coding [13].

The Golomb code was designed to encode non negative numbers with the assumption that the larger the number the less probability of its occurrence.

The Golomb code for numbers n depends on the choice of a parameter m. This parameter divides the number code in two parts, the q (quotient) and r (remainder). The fist step to construct the Golomb code is to calculate the q, the r and c as

$$q = \lfloor \frac{n}{m} \rfloor, r = n - qm, c = \lceil log(m) \rceil. \tag{3.6}$$

The quotient and the remainder are coded in different ways, q is codded in unary. Unary coding represents a natural number n with n ones followed by a zero. As an example, number 5 is represented as 111110. Zeros and ones are interchangeable as 5 can also be represented as 000001 but Golomb coding uses the first solution. The remainder is coded in truncated binary encoding. The first $2^c - m$ values of r are coded as unsigned integers, with $c - 1$ bits each, and the rest are coded in c bits each (ending with the biggest c-bit number, witch consists of c ones). The case when m is power of 2 ($m = 2^c$) is special because it only requires $log(m)$ bits. This special case has been conceived by Robert F. Rice and is often called Rice-Golomb coding [21], requiring less computation power as multiplications and divisions of 2 can be done in a much more efficient way. The case when $m = 1$ is also a special case, turning Golomb coding in simple unary coding.

As an example of the truncated binary encoding, with $m = 3$ produces $c = 2$ and three possible remainders 0, 1, and 2. Calculating $2^2 - 3 = 1$, so the first remainder is coded in $c - 1 = 1$ bit to become 0, and the remaining two are coded in two bits 10 and 11 respectively. An m of 5 results in $c = 3$, and five remainders (0 to 4). The first three are coded in 2 bits

and the other two in 3 bits. Thus, 00, 01, 10, 110 and 111. Table 3.4 shows Golomb code for $m = 2$ for values up to 10.

n	q	r	q code	r code	codeword
0	0	0	0	0	00
1	0	1	0	1	01
2	1	0	10	0	100
3	1	1	10	1	101
4	2	0	110	0	1100
5	2	1	110	1	1101
6	3	0	1110	0	11100
7	3	1	1110	1	11101
8	4	0	11110	0	111100
9	4	1	11110	1	111101
10	5	0	111110	0	1111100

Table 3.4: Golomb code for $m = 2$ (adapted from [13]).

Golomb coding is used in several lossless audio codecs, such as Sorten, Flac, Apple Lossless and MPEG-4, as well as in lossless video codecs such as JPEG-LS. Rice-Golomb coding is a very interesting solution to our task, specifically to encode oligonucleotide identifiers when the number of holes is small, or to encode the oligonucleotide occurrences when we are using high N orders as the number of occurrences for each oligonucleotide identifier is very small.

3.4 Solution adopted

After studying various compression methods and with time efficiency in mind, we have chosen to implement the max size compression method as well as Rice-Golomb coding. We are expecting low oligonucleotides counts for high oligonucleotide orders and high oligonucleotide counts for low oligonucleotide orders. As for oligonucleotide identifiers, when we are using hash tables, we expect values to be high because when we use hash tables there are considerable number of holes. For high values, we expect the max size compression to be better than Rice-Golomb coding, and Rice-Golomb coding better for low values. In order for the decoder to be able to reconstruct the original values, the header file of the compressed file contains (figure 3.4):

- Oligonucleotide order N.

- The number of total oligonucleotide count.
- Data structure type information, indicating what kind of data structure it is (simple array or hash table).
- In the case of simple array data structure, only the oligonucleotide counts are stored, so we store information identifying the compression method used (max size compression or Rice-Golomb coding), as well as the compression method properties (the m used for the case of Rice-Golomb coding, or the number of bits required to represent the biggest number in the case of max size compression).
- In the case of hash tables, we have to store both the oligonucleotide identifiers, as well as the oligonucleotide counts. Because of that, we store the compression method identifier as well as the compression method properties for both.

The sequence of encoded oligonucleotide identifiers do not require the signal to be stored, because all encoded values are positive, but for the oligonucleotide counts we need to store signal information, because of the mean predictor.

Table 3.5 shows compression results for *Ecoli bacteria* genome when using simple array are shown for oligonucleotide order of 1 to 11. As expected, for low oligonucleotide orders the max size compression produces much better results than Rice-Golomb coding. The only oligonucleotide order where max size compression, does not compress the values is for 1. This is because of the size the file header. Rice-Golomb coding only starts to compress values at oligonucleotide order of 8. Results could be better if we tested with higher m parameters, but for low oligonucleotide orders we already expected it to not perform well. The m used was 2.

Table 3.6 shows compression results for *Ecoli bacteria* genome ologonucleotide counts for orders of 12 to 15. Hash table is used for these orders when processing *Ecoli bacteria* genome. These values show that Rice-Golomb performed better than max size value compression. At these oligonucleotide orders the number of bits required to encode values using max size compression did not change, while the Rice-Golomb coding compression ratio decreased, showing that the majority of values of the oligonucleotide counts decreases for higher oligonucleotide orders. It also indicates that there are some oligonucleotide counts that stay high, resulting in constant max size data compression rate.

Table 3.7 shows compression results for *Ecoli bacteria* genome for ologonucleotide identifiers of 12 to 15. The hash table data structure is used for these orders when processing *Ecoli bacteria* genome. Analysing these values we can see that the higher the oligonucleotide order, the worse the compression results are. This happens because the higher the oligonucleotide order, the higher the number of holes is going to be, resulting in higher values. Comparing both

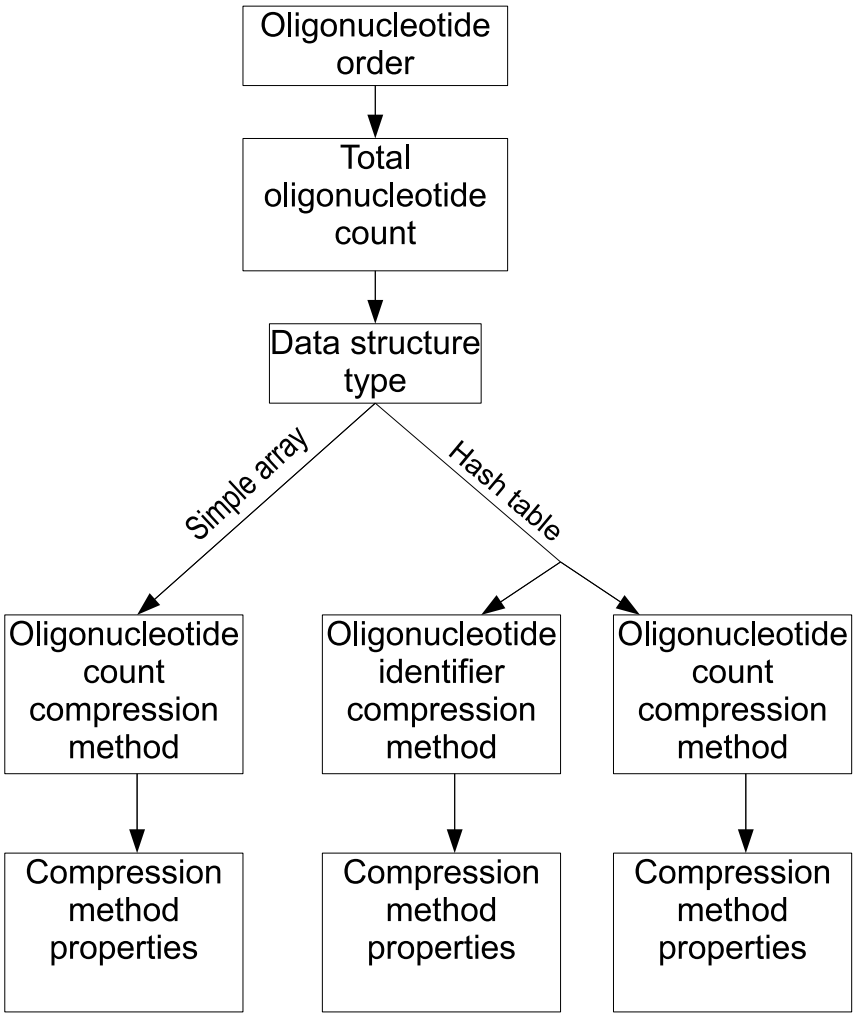

Figure 3.4: Compressed file header.

N	Uncompressed size (bytes)	Max Size compressed size (bytes)	Golomb Compression size (bytes)	Max Size compression ratio	Golomb compression ratio
1	16	25	4598	1.56	287.38
2	64	53	17880	0.83	279.38
3	256	153	64653	0.60	252.55
4	1024	529	87677	0.52	85.622
5	4096	1937	110075	0.47	26.88
6	16384	7185	129649	0.43	7.913
7	65536	24583	152696	0.37	2.32
8	262144	90129	187568	0.34	0.72
9	1048576	327697	274503	0.31	0.26
10	4194304	1179665	577075	0.28	0.13
11	16777216	4194321	1679653	25	0.10

Table 3.5: Simple array data compression results for *Ecoli bacteria* genome for oligonucleotide orders 1 to 11.

N	Uncompressed size (bytes)	Max Size compressed size (bytes)	Golomb Compression size (bytes)	Max Size compression ratio	Golomb compression ratio
12	13915692	3044058	1416695	0.219	0.101
13	16681292	3649033	1613280	0.219	0.097
14	17728768	3878168	1683278	0.219	0.095
15	18070484	3952918	1705390	0.219	0.094

Table 3.6: Hash table data compression results for oligonucleotide occurrences processing *Ecoli bacteria* genome for oligonucleotide orders 12 to 15.

compression methods we clearly see that max size compression performs better as expected.

Having analysed these results we can conclude that while using simple arrays, the best solution is the max data compression method. When using hash tables, the oligonucleotide counts should be encoded using Rice-Golomb encoding, while the oligonucleotide identifiers should be encoded using max data compression.

N	Uncompressed size (bytes)	Max Size compressed size (bytes)	Golomb Compression size (bytes)	Max Size compression ratio	Golomb compression ratio
12	27831384	2174343	1614181	0.078	0.059
13	33362584	3649049	4853994	0.109	0.145
14	35457536	4986233	17471851	0.140	0.493
15	36140968	6211745	67814733	0.172	1.877

Table 3.7: Simple array data compression results for oligonucleotide identifiers processing *Ecoli bacteria* genome for oligonucleotide orders 12 to 15.

Table 3.8 shows time efficiency when processing *Homo Sapiens* complete genome. Looking at the results, we see that the time it takes to load the results is much lower than processing the file. Results show the benefit of saving the information of previous analysis, greatly minimizing the runtime of future ones.

In this work we use the following definition of time efficiency:

$$\frac{\text{loading time}}{\text{processing time}} \tag{3.7}$$

N	Processing time (ms)	Loading time (ms)	Time efficiency
1	134677	0	0.0
2	173897	0	0.0
3	222402	0	0.0
4	281246	1	3.5556062E-6
5	347564	1	2.8771677E-6
6	387374	6	0.000015488907
7	451796	16	0.000035414214
8	495688	29	0.000058504543
9	611985	45	0.000073531214
10	836334	76	0.000090872785
11	1034276	182	0.0001759685
12	1149269	566	0.000492487
13	1361182	2161	0.0015875907
14	1523071	8913	0.0058519924
15	1702033	36824	0.021635303

Table 3.8: Results for time efficiency processing *Homo Sapiens* complete genome.

Chapter 4

Symmetry measures

In this chapter we will explore previously used mathematical methods to measure DNA symmetry such as skews, KullBack-Leibler divergence, Pearson's linear correlation coefficient and L^1 distance.

We also present a now approach based on equivalence tests, that we called equivalent pairs.

4.1 Skew of mononucleotide frequencies

When measuring mononucleotide frequencies, the most commonly used symmetry measure is the skew of mononucleotide frequecies.

Given a DNA sequence, the skews of mononucleotide frequencies are given by [24]:

$$ATskew = \frac{f_A - f_T}{f_A + f_T} \tag{4.1}$$

$$GCskew = \frac{f_G - f_C}{f_G + f_C} \tag{4.2}$$

where f_X denotes the observed frequency of the nucleotide X.

These values measure the frequency difference of two nucleotides and range between -1 and 1. Perfect symmetry occurs when the skew equals 0.

With this symmetry measure we can only compute values between two different mononucleotides, not a global symmetry measure. The AT and GC skews can be used to measure inverted symmetry as well as complement symmetry. Reverse symmetry cannot be measured because the reverse of a mononucleotide is always itself.

Figure 4.1 shows an example of mononucleotide skew results for *Saccharomyces cerevisiae* individual chromosomes.

	size	A (%)	T (%)	AT skew (%)	G (%)	C (%)	GC skew (%)
Chr. 1	230 203	30.33	30.39	−0.10	19.88	19.39	1.24
Chr. 2	813 140	30.70	30.95	−0.41	18.99	19.36	−0.97
Chr. 3	315 339	31.14	30.30	1.37	18.85	19.70	−2.21
Chr. 4	1 531 929	31.12	30.97	0.24	19.02	18.89	0.35
Chr. 5	576 870	30.60	30.89	−0.47	19.47	19.04	1.12
Chr. 6	270 148	30.70	30.57	0.21	19.41	19.32	0.22
Chr. 7	1 090 936	31.01	30.93	0.14	19.02	19.04	−0.08
Chr. 8	562 638	30.93	30.58	0.57	19.10	19.39	−0.74
Chr. 9	439 885	30.54	30.56	−0.03	19.47	19.43	0.12
Chr. 10	745 440	31.00	30.63	0.60	19.29	19.08	0.56
Chr. 11	666 445	30.92	31.01	−0.15	18.91	19.16	−0.67
Chr. 12	1 078 172	30.66	30.86	−0.33	19.21	19.27	−0.17
Chr. 13	924 430	30.97	30.83	0.23	19.09	19.12	−0.09
Chr. 14	784 328	30.80	30.56	0.38	19.30	19.34	−0.09
Chr. 15	1 091 283	31.10	30.74	0.58	19.01	19.15	−0.39
Chr. 16	948 061	31.01	30.93	0.12	19.04	19.02	0.04
Chr. mt	85 779	42.17	40.73	1.74	9.11	8.00	6.47

Figure 4.1: Single-stranded based composition (%) *Saccharomyces cerevisiae* individual chromosomes. The corresponding AT and GC skews. (from [4])

4.2 Relative abundance

The relative abundance is not by itself a symmetry measure. It computes a ratio between specific oligonucleotide frequency and the frequency of the nucleotide bases. The relative abundance of a dinuclotide can be measured by [22]:

$$\rho XY = \frac{f_{XY}}{f_X f_Y} \tag{4.3}$$

where f_{XY} denotes the observed frequency of dinucleotide XY and f_X denotes the observed frequency of nucleotide X.

When the relative abundance is higher than one it means that the given XY dinucleotide is over-represented in a sequence. If it is lower than one then the XY dinucleotine is under-represented.

By comparing oligonucleotide relative abundance with their conjugates, we can infer symmetry. For example when looking for reverse symmetry, if we calculate the relative abundance of AT and TA and their values are equal, then we can conclude the existance of symmetry. The closer the relative abundance values of the conjugates are, the more symmetry is present. If the nucleotide constituents of the oligonucleotide and its conjugate are different, then it is not possible to evaluate symmetry with the relative abundance.

Figure 4.2 shows an example of relative abundance for 16 different dinucleotides.

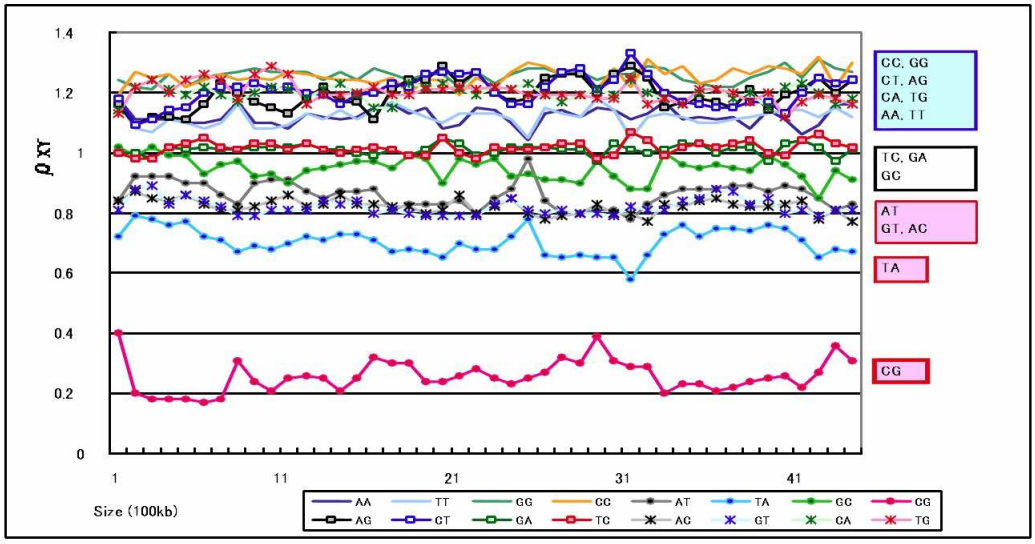

Figure 4.2: Relative abundance of 16 dinucleotides (from [22])

4.3 KullBack-Leibler divergence

The KullBack-Leibler divergence measures the difference between two probability distributions. For probability distributions P and Q of discrete random variables, their KL divergence is defined by [16]:

$$KLD(P\|Q) = \sum_{i=1} P_i log\left(\frac{P_i}{Q_i}\right) \tag{4.4}$$

It is the average of the logarithmic difference between the probabilities P and Q, where the average is taken using the probabilities P.

Both P and Q have to sum to 1. If 0 $log(0)$ appears in the formula, it will be interpreted as 0. The KL divergence is not symmetric which means that if the position of P and Q are switched the result may not be the same, $KLD(P,Q)$ is greater or equal to 0 for all distributions P, Q and does not have a maximum value. When $KLD(P,Q)$ equals 0, P is equal to Q resulting in perfect symmetry.

4.4 L^1 distance

The L^1 distance measures the sum of the absolute values of the differences between two probability distributions. For probability distributions P and Q of a discrete random variable their L^1 distance is defined by [4]:

$$L^1 = 1 - \frac{\sum_i |P_i - Q_i|}{\sum_i |P_i| + |Q_i|} \tag{4.5}$$

This value ranges from 0 to 1, where 0 means no symmetry and 1 means perfect symmetry. Figure 4.3 shows L^1 results for *Saccharomyces cerevisiae* individual chromosomes.

4.5 Pearson's linear correlation coefficient

Pearson's correlation coefficient measures the covariance of the two variables divided by the product of their standard deviations which may be written for the probability distributions P and Q of length n, as [18]:

$$S^C = \frac{n\sum_i P_i Q_i - \sum_i P_i \sum_i Q_i}{\sqrt{n\sum_i P_i^2 - (\sum_i P_i)^2}\sqrt{n\sum_i Q_i^2 - (\sum_i Q_i)^2}} \tag{4.6}$$

The values of the Pearson's correlation coefficient range from -1 to 1 , but we have to be careful analysing this value because it is sensitive to the relative abundance (4.3) of the oligonucleotides. For example, sequences with widespread oligonuleotide distributions result in higher S^C values [4].

Figure 4.3 shows S^C results for *Saccharomyces cerevisiae* individual chromosomes.

	AA	TT	AC	GT	AG	CT	CA	TG	CC	GG	GA	TC	S^1	S^C
Chr. 1	10.4	10.5	5.4	5.6	5.9	5.7	6.6	6.8	4.0	4.1	6.3	6.1	98.75	0.9964
Chr. 2	10.7	10.9	5.3	5.2	5.8	5.9	6.6	6.4	4.0	3.9	6.1	6.3	98.93	0.9981
Chr. 3	10.9	10.3	5.6	5.2	5.7	5.8	6.8	6.3	4.1	3.8	6.1	6.3	97.33	0.9914
Chr. 4	10.9	10.9	5.2	5.2	5.9	5.8	6.5	6.5	3.8	3.8	6.3	6.2	99.61	0.9997
Chr. 5	10.6	10.8	5.2	5.4	5.8	5.8	6.4	6.6	3.9	4.0	6.2	6.1	98.86	0.9982
Chr. 6	10.7	10.6	5.3	5.3	5.8	5.9	6.4	6.6	4.0	4.0	6.3	6.2	99.37	0.9993
Chr. 7	10.9	10.9	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.8	6.2	6.2	99.63	0.9999
Chr. 8	10.9	10.6	5.4	5.3	5.8	5.8	6.6	6.5	4.0	3.9	6.2	6.2	99.04	0.9990
Chr. 9	10.6	10.6	5.3	5.4	5.9	5.9	6.5	6.5	4.0	4.0	6.2	6.2	99.80	0.9999
Chr. 10	10.9	10.6	5.3	5.3	5.9	5.8	6.5	6.5	3.8	3.9	6.3	6.2	99.17	0.9997
Chr. 11	10.9	10.9	5.2	5.2	5.8	5.9	6.5	6.4	3.9	3.8	6.2	6.3	99.45	0.9992
Chr. 12	10.7	10.8	5.2	5.3	5.9	5.9	6.5	6.5	4.0	3.9	6.3	6.3	99.55	0.9999
Chr. 13	10.9	10.8	5.3	5.3	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.68	1.0000
Chr. 14	10.7	10.6	5.4	5.3	5.9	5.8	6.5	6.5	4.0	3.9	6.3	6.2	99.52	0.9999
Chr. 15	11.0	10.8	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.26	0.9998
Chr. 16	10.9	10.9	5.2	5.2	5.9	5.9	6.4	6.4	3.9	3.8	6.3	6.2	99.79	1.0000
Chr. mt	16.0	14.7	2.1	2.6	3.0	2.5	2.1	2.3	2.3	2.6	3.0	2.6	94.28	0.9975

Figure 4.3: Single-stranded base composition (%). The corresponding $L^1(S^1)$ and S^C symmetry levels for *Saccharomyces cerevisiae* individual chromosomes (from [4]).

4.6 Proportion of equivalent pairs

Traditional hypothesis testing access differences between two proportions (π_1 and π_2):

$$H_0 : \pi_1 - \pi_2 = 0 \text{ vs } H_1 : \pi_1 - \pi_2 \neq 0 \tag{4.7}$$

Equivalence tests aim to establish/reject the equivalence between the proportions:

$$H_0 : |\pi_1 - \pi_2| \geq \delta^{'} \text{ vs } H_1 : |\pi_1 - \pi_2| \leq \delta^{'} \tag{4.8}$$

In practice, we propose that the tolerance value ($\delta^{'}$) used in equivalence tests, depends on the magnitude of the two proportions in comparison $\delta^{'} = \frac{P_i+Q_i}{2}\delta$. Previous work[1] shows that a significance level δ of 0.05 is reasonable when processing DNA sequences.

For equivalence testing we compare the confidence interval of the proportion difference with a neighbourhood centred at zero, $V_{\delta'}(0)$. To the number of pairs for which the interval

$$\left[P_i - Q_i - 1.96\sqrt{\frac{P_i(1-P_i)+Q_i(1-Q_i)}{n}}; P_i - Q_i + 1.96\sqrt{\frac{P_i(1-P_i)+Q_i(1-Q_i)}{n}}\right] \tag{4.9}$$

falls within the tolerance interval

$$\left[-\frac{P_i+Q_i}{2}0.05; \frac{P_i+Q_i}{2}0.05\right] \tag{4.10}$$

we call equivalent pairs.

The proportion of equivalent pairs is:

$$\frac{\text{equivalent pairs}}{n} \tag{4.11}$$

The proportion of equivalent pairs ranges from 0 to 1. Where 0 means no symmetry and 1 means perfect symmetry.

4.7 Remarks

These symmetry measures are calculated after processing the DNA sequences. The oligonucleotide order distribution P contains all oligonucleotide frequencies present on the data structure, and distribution Q varies depending on which symmetry we are calculating (inverse, reverse, complement).

Half of the distributions values will be identical, which can increase or not the symmetry if they are taken into account. For example, when looking for dinucleotide inverted symmetry, the dinucleotide frequencies of AA and TT will be taken into account twice, comparing AA with TT and TT with AA. In this case, we decided to measure full distribution symmetry.

Also, when the conjugate of an oligonucleotide is itself, symmetry measures are afected if such frequencies are taken into account. This happens because each symmetry type contain different number of conjugates that are equal to themselves. For example, the inverse of TA is TA. The number of equal conjugates for the different types of symmetry is:

- Complement - 0.
- Reverse - 2^N when the oligonucleotide N order is an even number and 2^{N+1} when N is an odd number.
- Inverse - 0 when the oligonucleotide N order is an odd number and 2^N when N is an even number.

In these special cases, we opted to not include them in our measures, as they could increase symmetry levels if taken into account.

In order to have a wide scope into DNA symmetry we decided to implement all four measures. KullBack-Leibler divergence, L^1 distance, Pearson's linear correlation coefficient and our equivalent pair solution.

Chapter 5

Experimental results

In this chapter we present our experimental results obtained from analysing a selection of genomes. 48 complete genomes sequences were analysed for N oligonucleotide order up to 15. The selected genomes include archaea, bacteria, protozoa, fungi, insects, mammals, nematodes, fish and plants. Also, the 24 *Homo sapiens* individual chromosomes were analysed. These results can be found in appendix (A.2).

The selected genomes can be fount at NCBI (*ftp://ftp.ncbi.nih.gov/*), the full list follows: *Aeropyrum pernix*, *Haloarcula marismortu*, *Halobacterium salinarum R1*, *Methanococcus jannaschii*, *Pyrococcus horikoshi*, *Thermococcus kodakarensis*, *Bacillus anthracis Ames*, *Bacillus subtilis*, *Chlamydia trachomatis*, *Clostridium botulinum*, *Desulfovibrio vulgaris*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Pseudomonas aeruginosa*, *Pyrococcus furiosus*, *Staphylococcus aureus*, *Streptococcus mutans*, *Streptococcus pneumoniae*, *Candida albicans*, *Neurospora crassa*, *Apis mellifera*, *Drosophila melanogaster*, *Bos taurus*, *Canis familiaris*, *Equus caballus*, *Homo sapiens*, *Mus musculus*, *Macaca mulatta*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Pan troglodytes*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Gallus gallus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Danio rerio*, *Fugu*, *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Vitis vinifera*, *Dictyostelium discoideum*, *Leishmania infantum*, *Plasmodium falciparum*, *Trypanosoma brucei*.

Additionally, the complete *Homo sapiens* genome was analysed up to order 18 and its individual chromossome 1 was analysed up to order 30.

We will compare the results of the different kinds of symmetry, study the impact of very high N oligonucleotide orders and genome size on DNA symmetry as well to try to find correlation between genome types. Finally, the different measures will be evaluated in regard to their effectiveness as symmetry measures. Plots presented were created using jfreechart package library for java, ibm spss as well as microsoft office.

5.1 Types of symmetry

In order to visually assess DNA symmetry we have chosen to plot the frequencies of the oligonucleotides and their conjugates. The closer the points are to $y = x$ the more symmetric the sequence is. When looking at the *Homo sapiens* complete genome plot for inverted symmetry (figure 5.1) we clearly see that for N orders up to 9, all points are very close to $y = x$ confirming symmetry. Analysing the plots for the reverse (figure 5.2) and complement (figure 5.3) conjugates we can visually confirm that reverse and complement symmetries and not present at any level.

Furthermore we can see that the reverse and complement plots look very similar between then, resulting in similar low levels of symmetry at the same oligonucleotide orders. This may happen as a consequence of inverted symmetry. When inverted symmetry is present (i.e the frequency of the oligonucleotides equals the frequency of their inverse conjugates), the frequency of the reverse and complement conjugates are equal between themselves. For example, if $f_{AC} = f_{GT}$ and $f_{CA} = f_{TG}$, then the contribution, to the scatter plots, of the reverse pair (f_{AC}, f_{CA}) and that of the complement pair (f_{AC}, f_{TG}) are the same. Previous studies [15] have not shown this similarity between reverse and complementary symmetry. Most probably because the oligonucleotides that are the same as their conjugates are taken into account. As we have seen before the number of these occurrences depend on the type of symmetry analysed. This number is 0 for complement symmetry but for reverse this is not the case, as we seen before (chapter 4.7).

5.2 Large oligonucleotides

In order to determinate at witch order does inverted symmetry stop for the human genome, the complete genome was analysed up to order 18. After verifying, that for order 18 symmetry was not present, and because of memory and time constraints, we decided to process only the first chromosome to orders up to 30 to better evaluate the symmetry measures. Looking at (figure 5.4) we can see the behaviour of all four symmetry measurements, L^1 distance, KullBack-Leibler divergence, Pearson's linear correlation as well as our own measure (proportion of equivalent pairs). We can see that L^1 distance and KullBack-Leibler behave in the same manner, as both values tell us that around oligonucleotide order of 13, symmetry starts to greatly decrease. Pearson's linear correlation on the other hand still reveals high symmetry levels even at oligonucleotide order of 18. The Equivalent proportion is the most demanding, as it tells us that symmetry starts to decrease much sooner (at olygonucleotide order of 7).

Looking at results for the individual chromossome number 1 (figure 5.5), we see a simi-

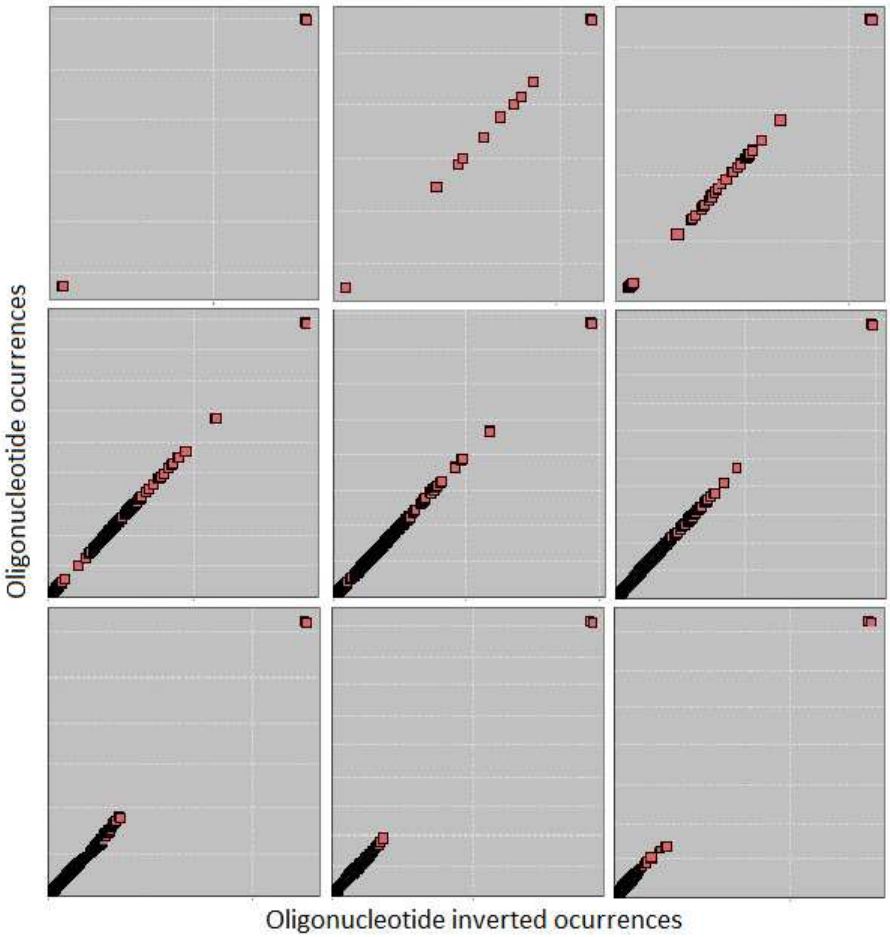

Figure 5.1: Frequency of oligonucleotides (y-axis) versus frequency of their inverse (x-axis), for *Homo sapiens* complete genome, oligonucleotide order from 1 to 9.

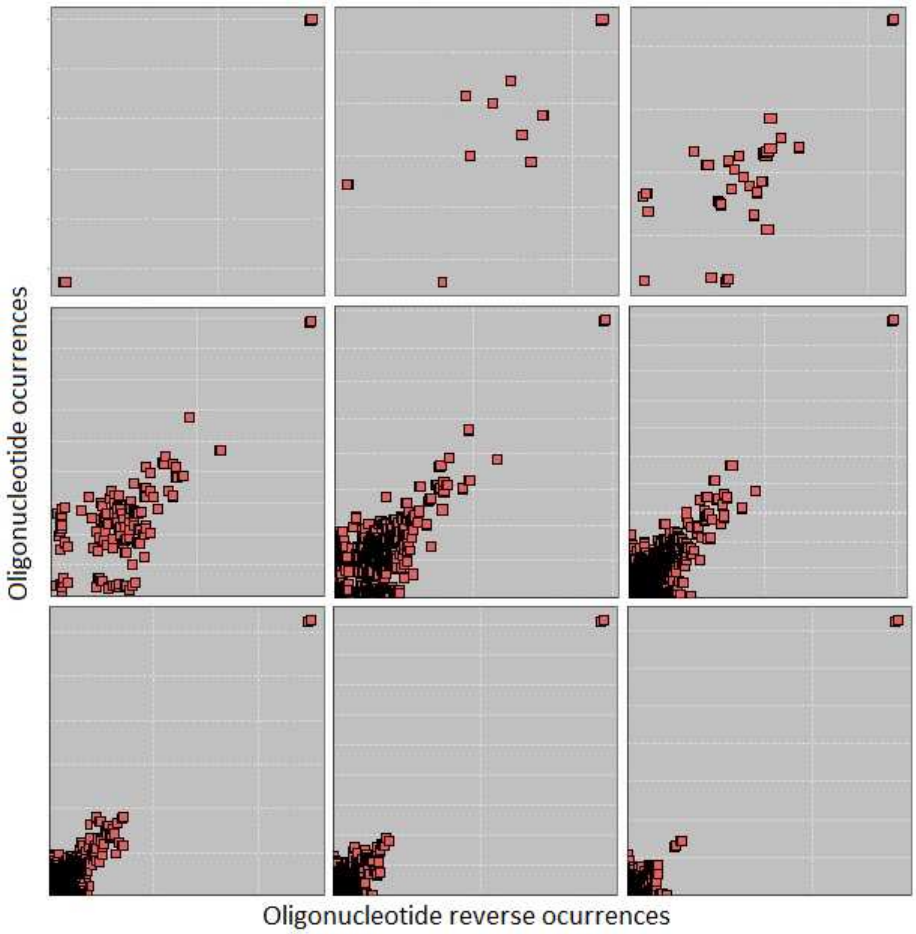

Figure 5.2: Frequency of oligonucleotides (y-axis) versus frequency of their reverse (x-axis), for *Homo sapiens* complete genome, oligonucleotide order from 1 to 9.

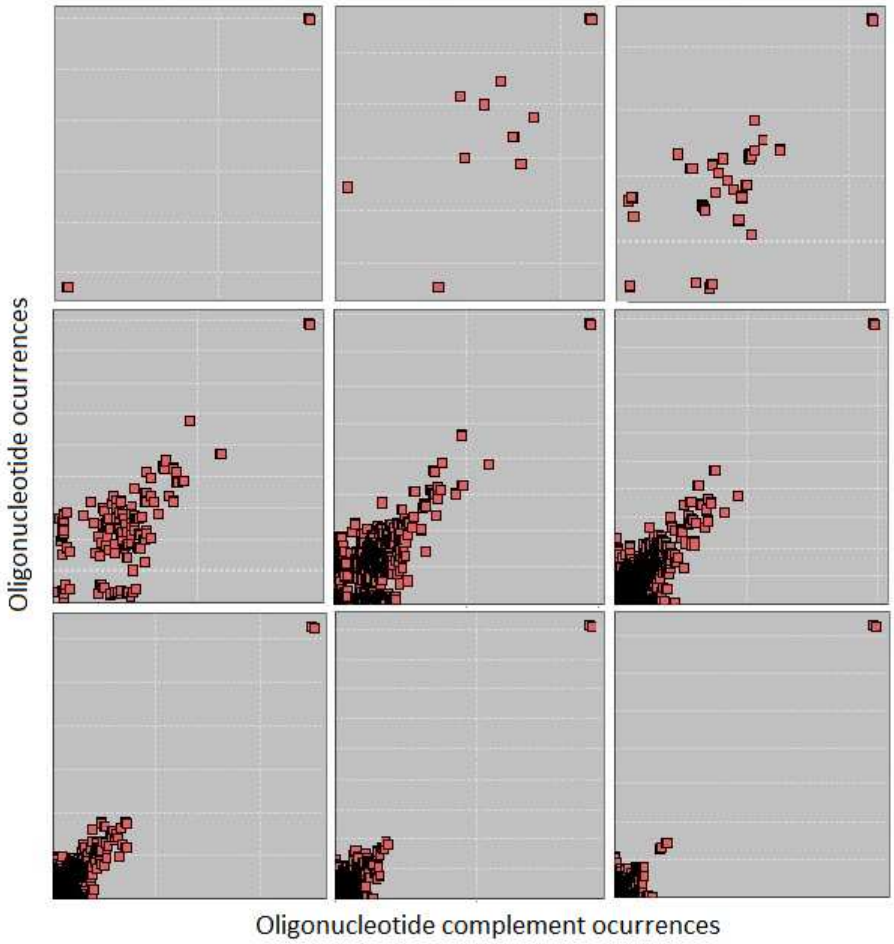

Figure 5.3: Frequency of oligonucleotides (y-axis) versus frequency of their complement (x-axis) for *Homo sapiens* complete genome, oligonucleotide order from 1 to 9.

lar behaviour of the measurements, but the symmetry starts to decrease sooner, at around oligonucleotide order of 11 for L^1 distance and KullBack-Leibler, and around oligonucleotide order of 6 for the Equivalent pair proportion. Pearson's linear correlation coefficient continues to give very high values, even at oligonucleotide order of 30. These results confirm the global assumption that the bigger the DNA sequence, the higher the symmetry level is and that the higher the oligonucleotide order, the lower the symmetry found.

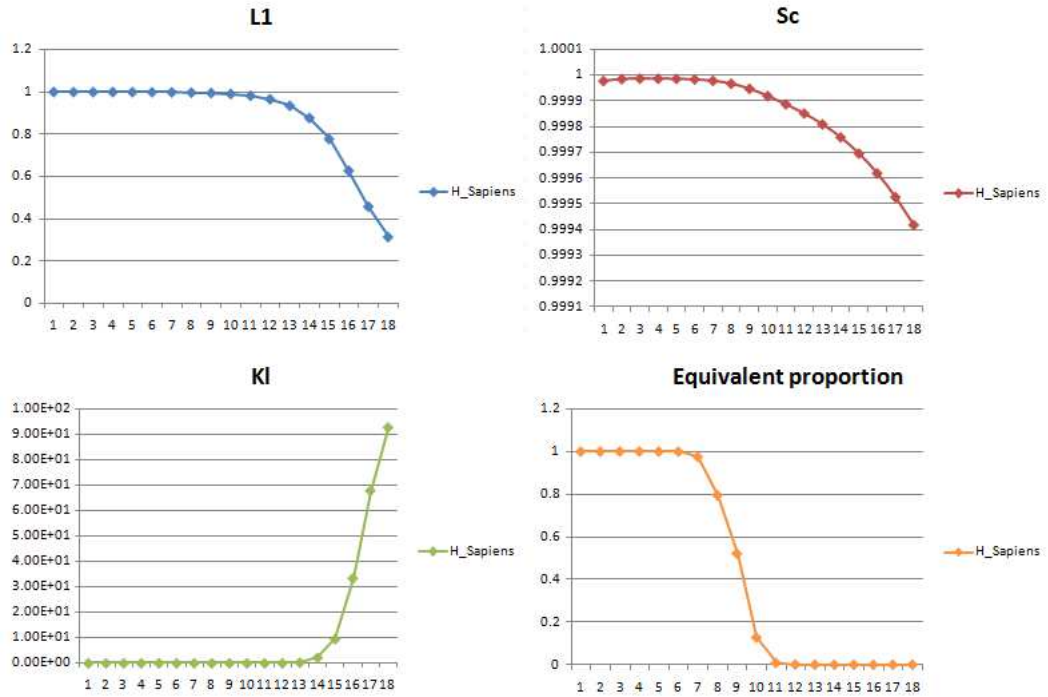

Figure 5.4: *Homo sapiens* complete genome symmetry measures for oligonucleotide order up to 18.

5.3 Genome size impact on symmetry

Looking at figures 5.6, 5.7 and 5.8 we see several genome results and respective symmetry measurements. Measurements reflect that the bigger the genome size is, the more symmetric it is. Small genomes such as *Escherichia Coli* result in weak symmetry levels at low oligonucleotide orders, while large genomes such as the case of *Homo Sapiens* or other mammals keep high symmetry levels on higher oligonucleotide order.

Figure 5.9 and 5.10 show the whiskers boxplot and the interval plot for various genome sizes. The measurements seem to indicate that higher genome sizes tend to show higher degrees of symmetry.

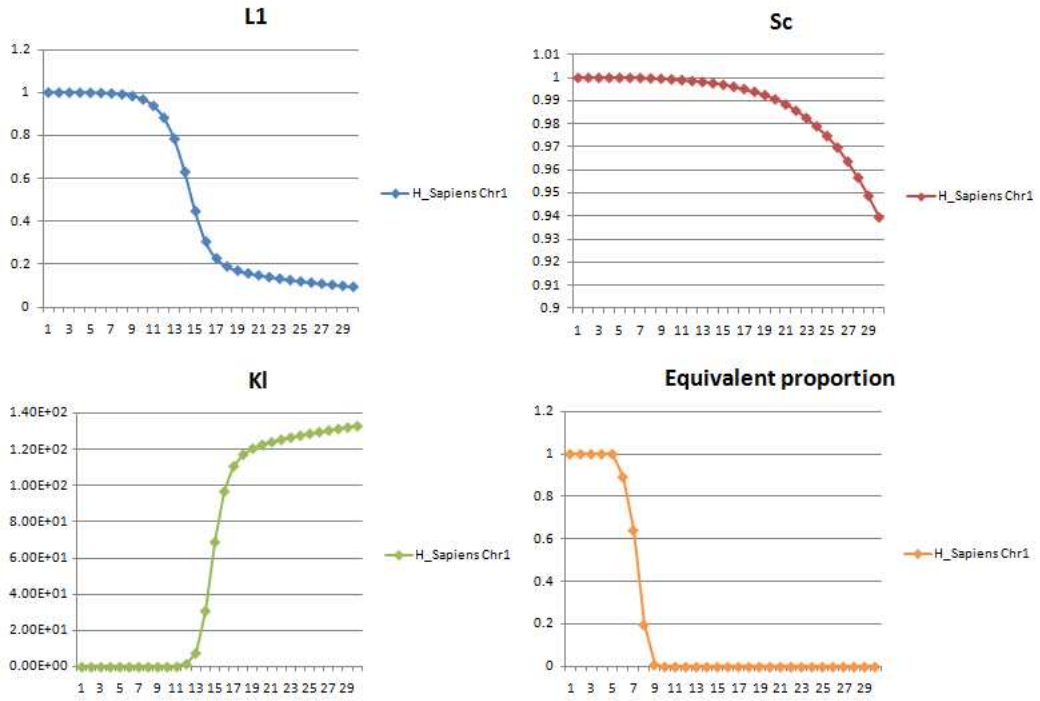

Figure 5.5: *Homo sapiens* chromosome 1 symmetry measures for oligonucleotide order up to 30.

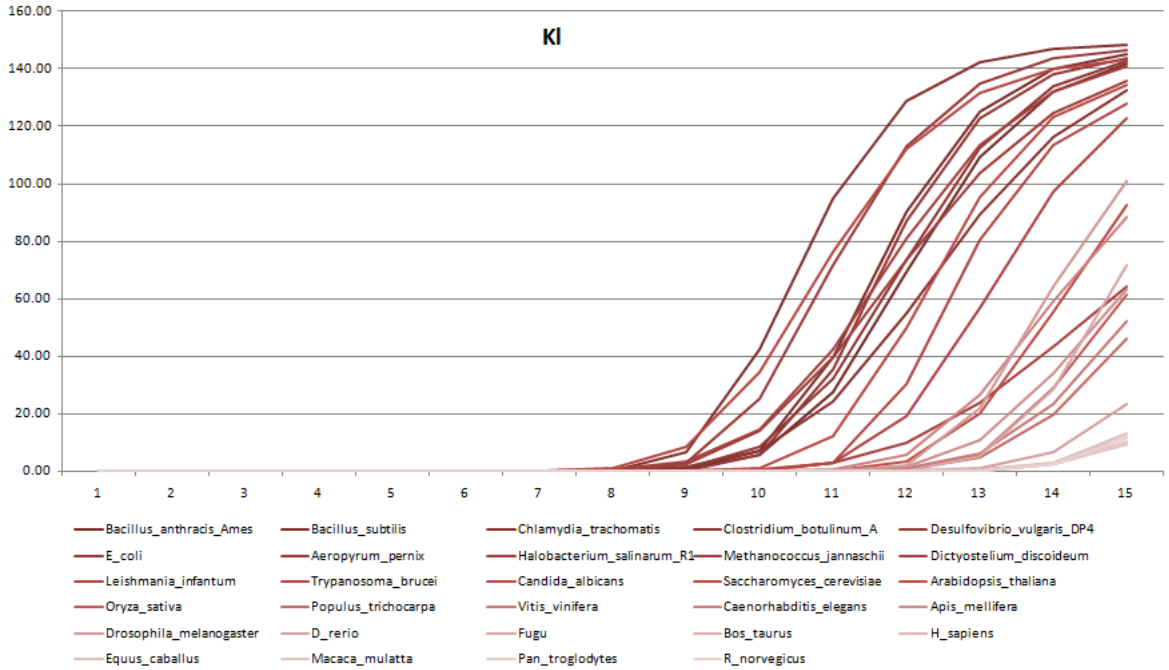

Figure 5.6: Several genomes, KullBack-Leibler divergence measurements for oligonucleotide order up to 15.

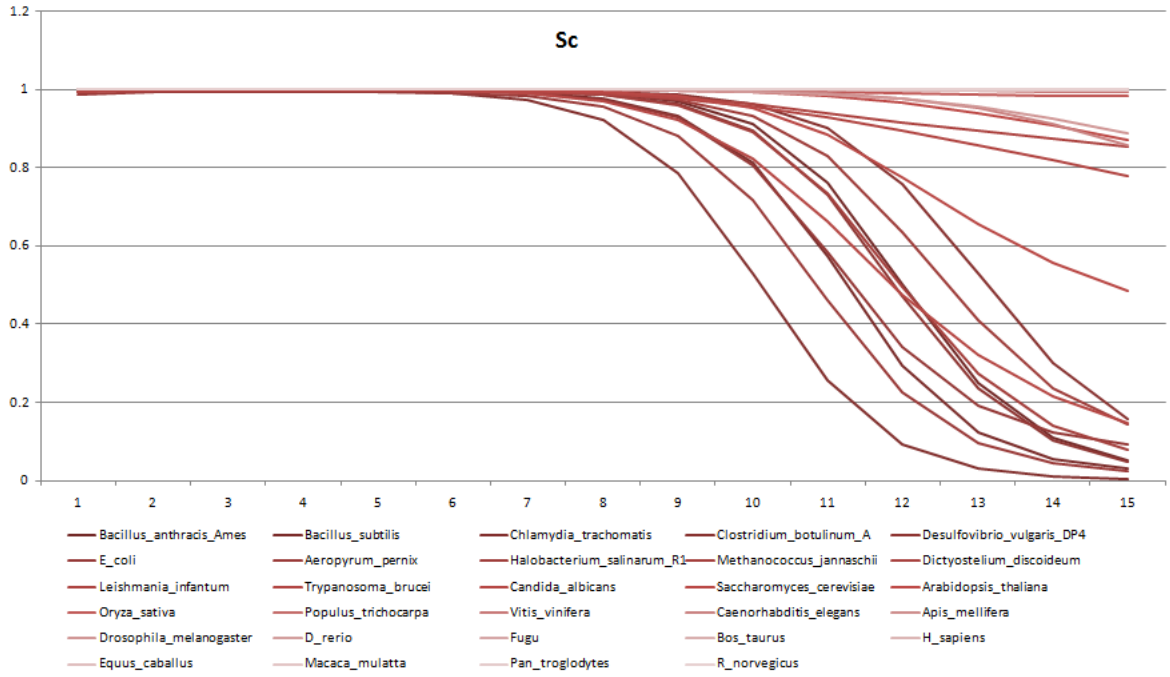

Figure 5.7: Several genomes, Pearson's linear correlation coefficient measurements for oligonucleotide orders up to 15.

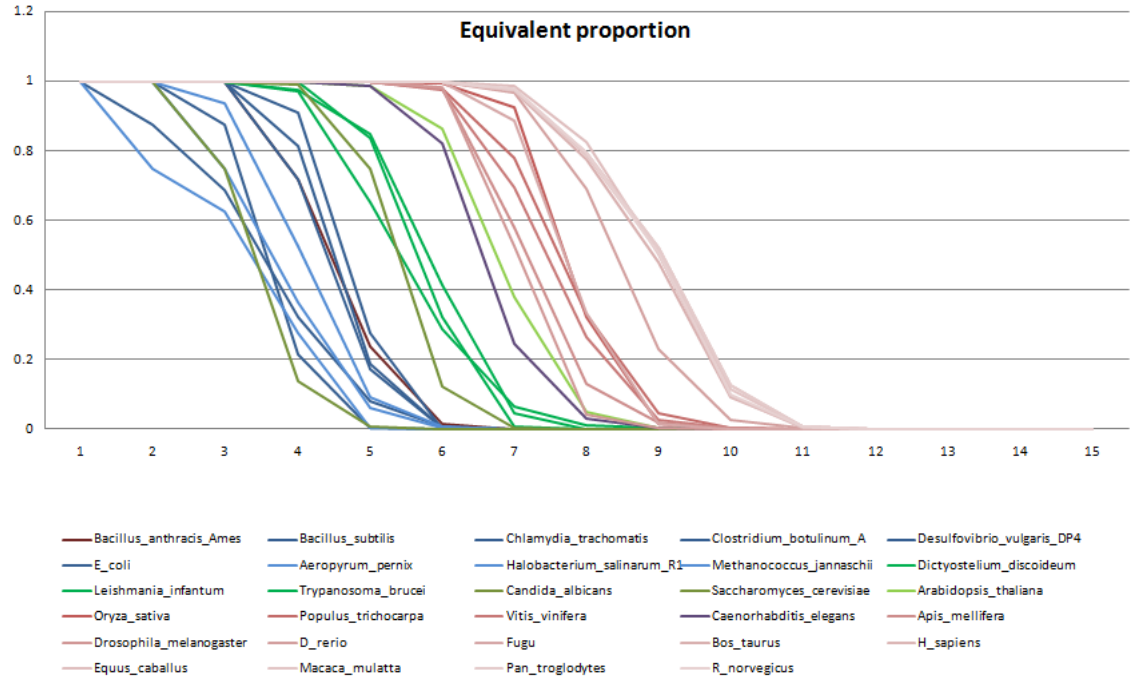

Figure 5.8: Several genomes, equivalent pair proportion orders up to 15.

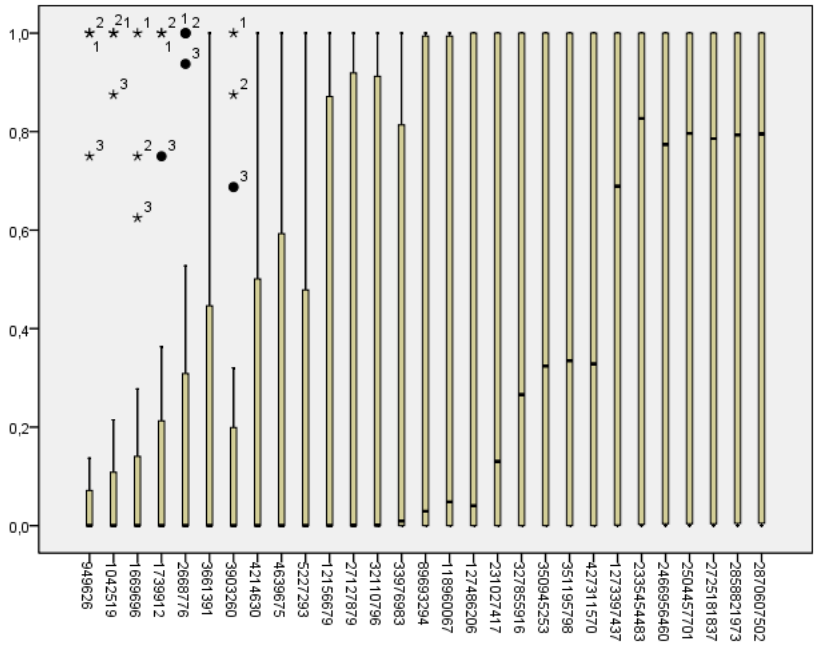

Figure 5.9: Whiskers boxplot for several genomes, equivalent pair proportion vs genome size (total nucleotide count).

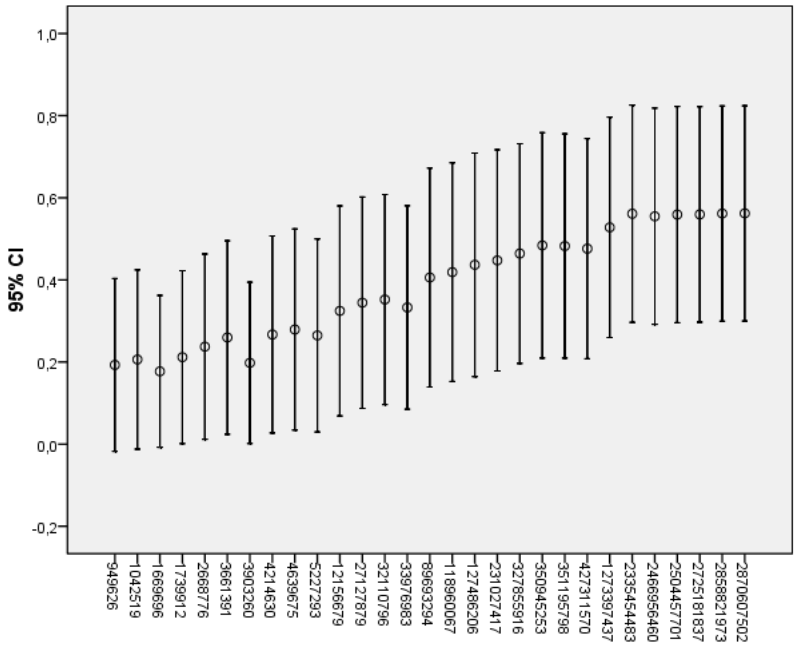

Figure 5.10: Interval plot for several genomes, Equivalent pair proportion vs genome size (total nucleotide count). Confidence interval of 95%.

5.4 Genome correlation

In order to try to find a relation between different genome types (archaea, bacteria, archaea, fungi, insects, mammals, nematodes, fish and plants) we calculated the mean of the symmetry measurements obtained from the genomes of similar organisms. The results were obtained for oligonucleotide orders up to 15 their respective symmetry results (figures 5.11,5.12,5.13) were plotted. Unfortunately results were not conclusive, because the organism types in which correlation was found, correspond exactly to the types that have similar genome sizes. As we can see in figure 5.12, bacteria, archaea and fungi have similar levels of symmetry for the same oligonucleotide orders, but they also have in common the small genome size. This only furthermore confirms the impact of the genome size on symmetry but does not allow us to make definitive conclusions in relation to the correlations.

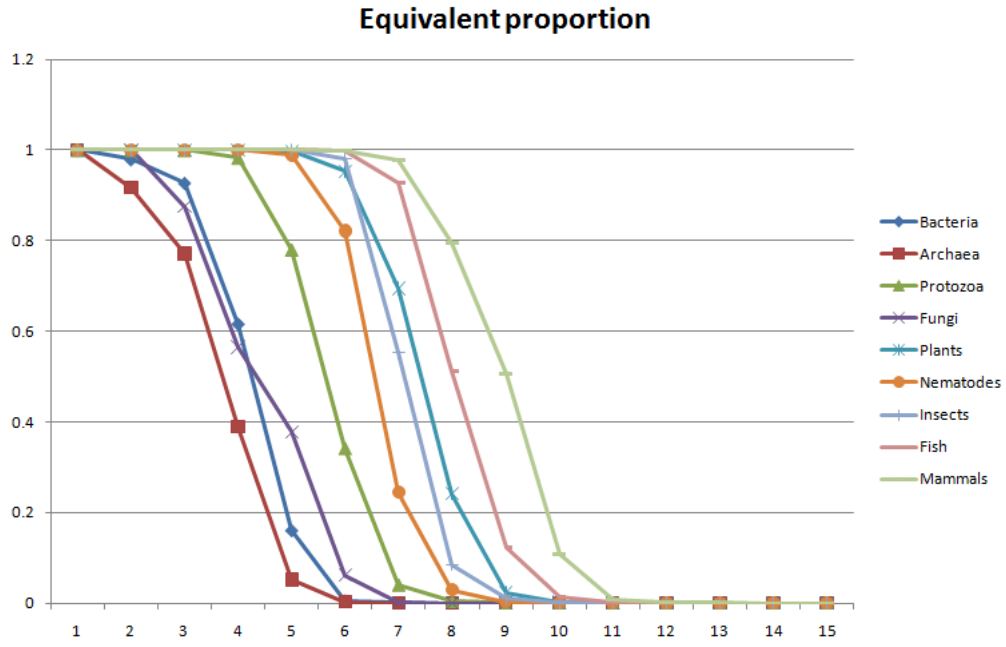

Figure 5.11: Equivalent proportion measurements for different genome types. Oligonucleotide order up to 15.

5.5 Measures evaluation

After analysing all of our results, we can conclude that our method (equivalent pairs) is a suitable symmetry measurement. Of all the methods implemented it is the most demanding as well, it reveals lack of symmetry at lower orders than the other measures. Looking at L^1

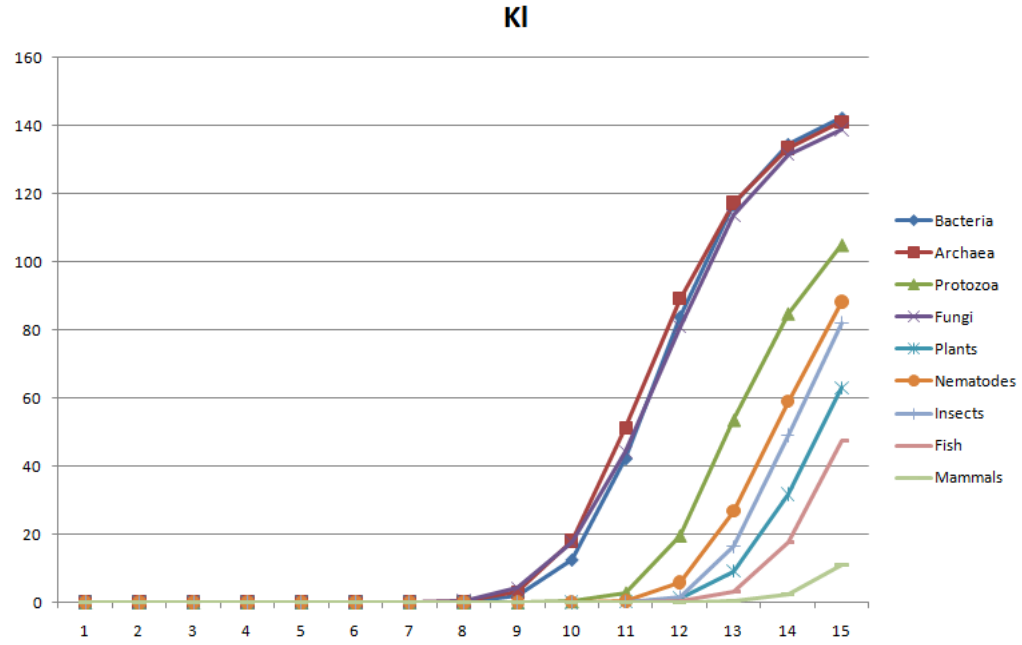

Figure 5.12: KullBack-Leibler divergence measurements for different genome types. Oligonucleotide order up to 15.

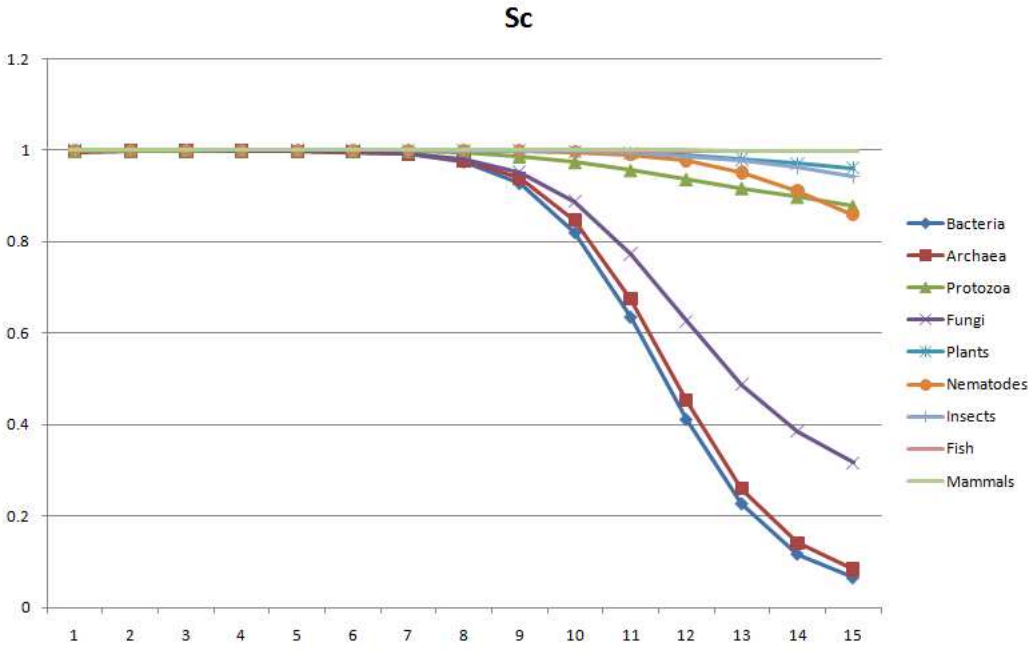

Figure 5.13: Pearson's linear correlation coefficient measurements for different genome types. Oligonucleotide order up to 15.

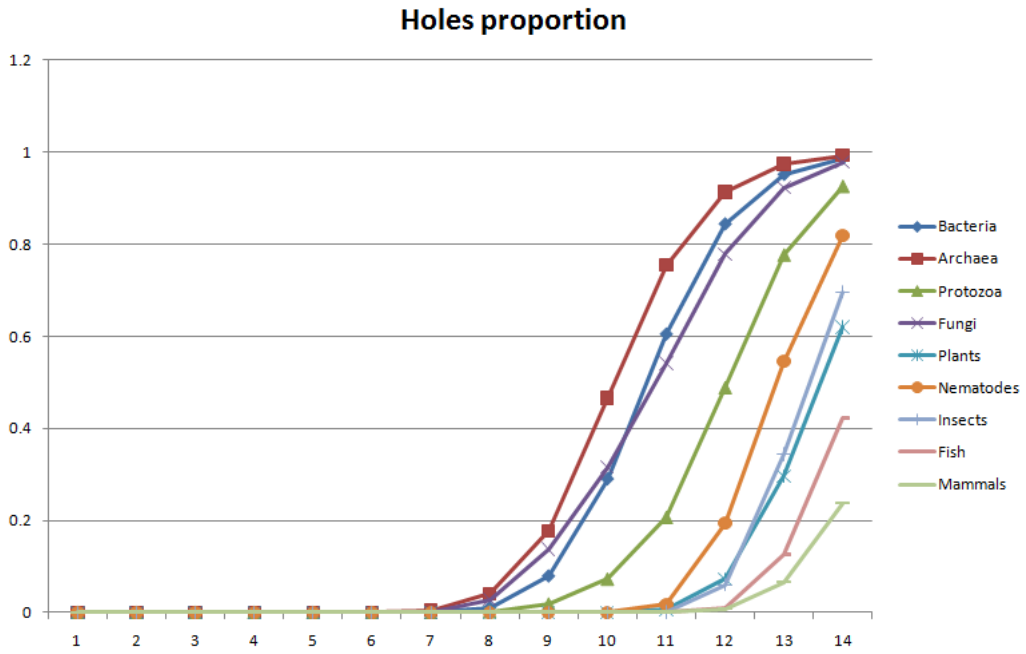

Figure 5.14: Hole proportion for different genome types. Oligonucleotide order up to 15.

distance and KullBack-Leibler divergence we can also conclude that they are also good symmetry measures and both return similar results for the same oligonucleotide order. On the other hand, Pearson's linear correlation proved not to be a good symmetry measure. Even at high oligonucleotide orders, results stay almost constant compared to low oligonucleotide orders. It has been documented before [4] that this correlation is sensitive to outliers (over-represented oligonucleotides). In figures 5.1, 5.2 and 5.3 we can see example of outliers for the *Homo sapiens* genome. There are a couple of points that are very far from the rest. Even so, results so far, indicated that Pearson's correlation coefficient resulted in similar values as L^1 distance. This is because higher oligonucleotide orders where not explored. As we can see from figure 5.15 for the case of the *Homo Sapiens* chromossome 1, at oligonucleotide orders higher than 10 L^1 distance and Pearson's correlation coefficient give very different results.

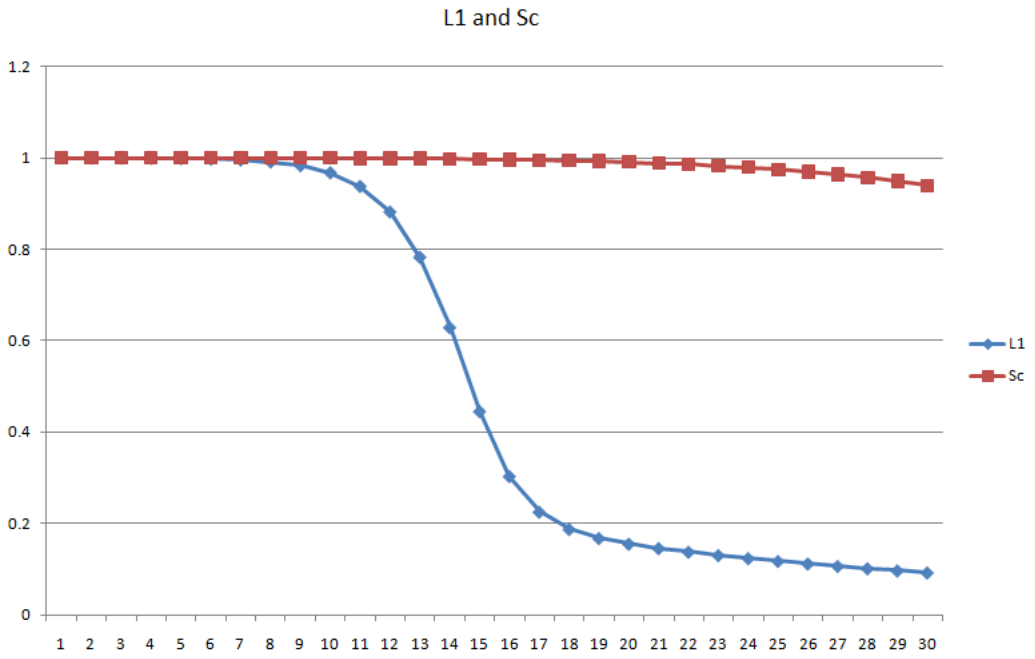

Figure 5.15: *Homo sapiens* chromossome 1, L^1 distance and Pearson's linear correlation coefficient.

Chapter 6

Conclusion and future work

In this dissertation we focused on measuring DNA symmetry. The main constatations were the correlation between genome size and symmetry levels, as well as the correlation between oligonucleotide orders and symmetry levels. The bigger the genome size is, the higher the symmetry level found. Also, the higher the oligonucleotide order, the lower the symmetry found.

We confirmed that reverse and complement symmetry are not present at any order, and that inverse symmetry is present in all genomes analysed at least at some oligonucleotide orders.

We found that for large genomes, symmetry is found past the oligonucleotide order of 10 previously documented in the literature.

Unfortunately we could not find any relation between genome types and respective symmetry, because results coincided with the impact of genome size, furthermore confirming the bigger the genome size, the higher the symmetry level found.

Symmetry measure wise we developed our own method (equivalent pairs) and concluded that it is suitable for measuring symmetry. Also L^1 distance and KullBack-Leibler divergence were confirmed as suitable measures, while Pearson's linear correlation is not very suitable as a symmetry measure.

Possible future work includes implementing a user friendly graphical interface, as well as processing coding and non coding parts of DNA separately.

Appendix A

Appendix

A.1 Implemented tool usage

Usage:
java -jar ReadingDNA.jar [OPTIONS1] FileFolderName Norder [OPTIONS2]

OPTIONS1:
[-s FileName1]
[-l FileName2]

OPTIONS2:
[-r](Reversed symmetry)
[-c](Complemented symmetry)
[-p](Create plot)

NOTES:
. FileName1 is the name of the file to save the compressed results
. FileName2 is the name of the file to load the compressed results
. FileFolderName is the name of the file or folder to process
. If neither -r or -c are chosen, by default inverted symmetry is calculated
. If OPTIONS1 is -s then FileFolderName argument isn't necessary

A.2 DNA sequence results

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1669696	4	0	0.9904449672275671	0.9882316008313337	2.88793777058321E-4	4
2	1669695	16	0	0.9893118204222927	0.9933719324695345	5.818526847955693E-4	12
3	1669694	64	0	0.9861232058089686	0.9943618402083675	9.528734643453228E-4	40
4	1669693	256	0	0.982954351488567	0.9949029039406069	0.0014206693617680774	71
5	1669692	1024	0	0.9763309640340854	0.9941370727288781	0.0024898070040726148	4
6	1669691	4096	0	0.9660691708825165	0.9912797765966631	0.005583723523345469	0
7	1669690	16384	0	0.9417125334642957	0.9820380823597412	0.017340155901552034	0
8	1669689	65536	137	0.8929028100442657	0.9553617562521795	0.1011461658058124	0
9	1669688	262144	20084	0.7973932854521324	0.8825589434005567	3.0292545431730575	0
10	1669687	1048576	401256	0.6287663496212164	0.7190439350445647	25.342302277923846	0
11	1669686	4194304	3072301	0.4012071730852388	0.4598770720127692	71.81039189079085	0
12	1669685	16777216	15340254	0.2021788540952335	0.22599767830463127	112.92316265125557	0
13	1669684	67108864	65528172	0.08589170166330873	0.09743748711521297	134.66879453187903	0
14	1669683	268435456	266800318	0.03435322752881831	0.04450680211201007	143.48611719982742	0
15	1669682	1073741824	1072087747	0.015038791817843178	0.025411606795818914	146.61303257268415	0

Table A.1: *Aeropyrum pernix*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	3131724	4	0	0.9994475886125341	0.9999878736951001	1.3544170891980712E-6	4
2	3131723	16	0	0.9987399907335355	0.9999630841953828	7.778821036974867E-6	16
3	3131722	64	0	0.9975661952114524	0.9999276529424234	2.9528365919661524E-5	64
4	3131721	256	0	0.9949673039201129	0.9998222118370328	1.2598534825570152E-4	192
5	3131720	1024	0	0.9899211934655716	0.999470013935842	5.084582148381395E-4	124
6	3131719	4096	0	0.9811129287142301	0.9985459093288344	0.001963630380004912	13
7	3131718	16384	0	0.9635490807282137	0.9960015966176328	0.007760206529430095	0
8	3131717	65536	164	0.9294118849180817	0.9883933906807888	0.05647831319025	0
9	3131716	262144	17363	0.86317533262914	0.9655301475944978	1.3267300115670249	0
10	3131715	1048576	331771	0.7434559019578729	0.9033235197978411	11.239640212224097	0
11	3131714	4194304	2718088	0.561896776014668	0.7551326545188028	40.11938851158076	0
12	3131713	16777216	14552133	0.35598440853296587	0.5014489401133503	81.12530409867394	0
13	3131712	67108864	64389088	0.18848093311262337	0.2534419078700521	115.00617244083863	0
14	3131711	268435456	265469852	0.08644188432457534	0.10424834782718764	134.3799038487688	0
15	3131710	1073741824	1070673848	0.03670263210833702	0.03993520402039508	143.13078397497574	0

Table A.2: *Haloarcula marismortu*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2668776	4	0	0.9982328977778577	0.9998741359137966	2.614327464431936E-5	4
2	2668771	16	0	0.9967539365498201	0.9998729742838341	5.330167558056474E-5	16
3	2668766	64	0	0.9948155814335164	0.999787599936451	1.1431078648887708E-4	60
4	2668761	256	0	0.9925403586158521	0.9996666035135615	2.70728981668999E-4	135
5	2668756	1024	0	0.9877126271566228	0.9993847431536405	7.798732098549989E-4	92
6	2668751	4096	0	0.9784311087845963	0.9985872672374384	0.002707605842769701	24
7	2668746	16384	0	0.9598905253628484	0.9966425009172685	0.010345149375638088	0
8	2668741	65536	683	0.9233844723036069	0.9910070462433751	0.14677950896401115	0
9	2668736	262144	36529	0.8534946881220173	0.9745445298894205	2.547403007533582	0
10	2668731	1048576	460938	0.735166639125487	0.9325766524459627	13.982418722626639	0
11	2668726	4194304	3073712	0.5701154783218659	0.8310722363106045	39.541435237346704	0
12	2668721	16777216	15115416	0.3915976979234622	0.6371432547690112	73.39073960088581	0
13	2668716	67108864	65042368	0.2396365892811374	0.41077079759147006	103.8977284089436	0
14	2668711	268435456	266128411	0.1352720470669173	0.23709454332181903	124.47871344111236	0
15	2668706	1073741824	1071312565	0.07428244250209648	0.14532129707100083	135.92554133878727	0

Table A.3: *Halobacterium salinarum*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1739912	4	0	0.9967940907356234	0.9998463335942206	4.1643581201005686E-5	4
2	1739909	16	0	0.9962687703782209	0.9998396162385692	1.1084092152880252E-4	16
3	1739906	64	0	0.9941318668939586	0.9997733467758994	2.2318694452055866E-4	48
4	1739903	256	0	0.9912644555472345	0.9996322780841193	5.514629268689762E-4	93
5	1739900	1024	0	0.9855014656014713	0.9992480958557437	0.001603814611654995	64
6	1739897	4096	2	0.9744973409345495	0.9981486941460149	0.005037604889317804	9
7	1739894	16384	275	0.9536259105439757	0.9950916338550662	0.04043102788709918	0
8	1739891	65536	7242	0.9139020777738376	0.9860773639709908	0.4711417438823661	0
9	1739888	262144	82916	0.8417254443964209	0.9597023117519168	3.2537177006108857	0
10	1739885	1048576	604059	0.7209097152972754	0.8899078004125692	14.524528674068241	0
11	1739882	4194304	3348828	0.547680819733752	0.7337306784856887	42.352225953853974	0
12	1739879	16777216	15538316	0.3553597692713114	0.494076950764949	81.17770145474641	0
13	1739876	67108864	65608101	0.1966875800344392	0.271996018371263	113.49526932753125	0
14	1739873	268435456	266802705	0.0985433994320275	0.1403915743037103	132.153177746344	0
15	1739870	1073741824	1072053308	0.04987039261554027	0.08071026353795381	140.6992292503194	0

Table A.4: *Methanococcus jannaschii*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1738505	4	0	0.9933845459173255	0.9955093250926719	2.0110216876156404E-4	4
2	1738504	16	0	0.9909646454652966	0.9971462904645912	3.7022596677069075E-4	16
3	1738503	64	0	0.987761309586466	0.9971317206249736	5.7823053767423E-4	46
4	1738502	256	0	0.985739446949155	0.996914813798101	8.851158237147562E-4	65
5	1738501	1024	0	0.980138636676079	0.9959175394754588	0.0017141815957393053	4
6	1738500	4096	0	0.9692838654012079	0.9925664058448512	0.004515515944491309	0
7	1738499	16384	4	0.94553520019281	0.9822415647727836	0.017090989511756898	0
8	1738498	65536	708	0.8982558507401216	0.9515196214419812	0.15936227541581957	0
9	1738497	262144	26863	0.8051748147969194	0.8650471726586295	2.8018717641959414	0
10	1738496	1048576	403468	0.6407187592033574	0.6721020940024806	22.274267394931332	0
11	1738495	4194304	3028417	0.41047572756896056	0.39846406889249153	69.4354210821322	0
12	1738494	16777216	15264487	0.19964636058565632	0.18229422629093517	113.82915740702916	0
13	1738493	67108864	65449098	0.07930201617147725	0.0771686920096679	136.07485772187133	0
14	1738492	268435456	266725453	0.03002717297519919	0.03644189385876597	144.2127964384632	0
15	1738491	1073741824	1072015755	0.013002080539962546	0.02071597005967017	146.87298614885722	0

Table A.5: *Pyrococcus horikoshii*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2088749	4	0	0.9971633738663669	0.9882167635798067	2.8051443144579476E-5	4
2	2088748	16	0	0.9968780341142158	0.998661230609546	6.102436022800638E-5	16
3	2088747	64	0	0.9946288372885754	0.9989736034674522	1.217080184598198E-4	64
4	2088746	256	0	0.9920282312928427	0.9986587047617858	2.888206706063073E-4	137
5	2088745	1024	0	0.9866154078166555	0.997525682411649	8.548773172316873E-4	28
6	2088744	4096	0	0.9757093257957893	0.9943235554680705	0.0029318800552274554	0
7	2088743	16384	0	0.9526227017876302	0.984616284834147	0.01143978168712286	0
8	2088742	65536	31	0.9072800757585188	0.956208877909526	0.05434622562031839	0
9	2088741	262144	10456	0.818584017836582	0.8797820485210437	1.580016972800673	0
10	2088740	1048576	315446	0.6586257743903023	0.7107938539881367	19.522019584807122	0
11	2088739	4194304	2841109	0.43171885046432323	0.4554903899269669	65.3507709400435	0
12	2088738	16777216	14998130	0.22342342601130438	0.2306276921950191	109.14359341538255	0
13	2088737	67108864	65136163	0.09888463698397643	0.10866520624347543	132.64377444063632	0
14	2088736	268435456	266390742	0.04353829301548884	0.057408574707570645	142.10087589835948	0
15	2088735	1073741824	1071672429	0.02254091591322016	0.037341074908452156	145.47684350100693	0

Table A.6: *Thermococcus kodakarensis*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	5227293	4	0	0.9965456307882493	0.9997107307610354	4.327178570526295E-5	4
2	5227292	16	0	0.9957748677517919	0.9997336855283857	8.704106016590967E-5	16
3	5227291	64	0	0.9934721445582425	0.9996014461003521	1.6627447973691166E-4	64
4	5227290	256	0	0.9919017311073233	0.9994769669659175	2.9715052203962007E-4	184
5	5227289	1024	0	0.9882158801627383	0.9991478162797465	6.388364751788644E-4	244
6	5227288	4096	0	0.981986835238464	0.9984506737525767	0.0016598958162715948	53
7	5227287	16384	0	0.9681775651499526	0.9964155233118569	0.005710771157420807	2
8	5227286	65536	30	0.9401524997866962	0.990026179495708	0.02444596174551268	0
9	5227285	262144	6924	0.8859245286989326	0.9701432006576197	0.41023039643172304	0
10	5227284	1048576	203554	0.7854331235876987	0.9115558981290521	5.4356384213664555	0
11	5227283	4194304	2169815	0.619959546862108	0.7628918284578239	27.55370194270095	0
12	5227282	16777216	13363255	0.4071511734013967	0.5028017188204253	69.19839586801281	0
13	5227281	67108864	62702489	0.21460832122856988	0.2500439610125919	109.07998849074852	0
14	5227280	268435456	263545550	0.09430621661743777	0.10889636689045032	132.08789801689326	0
15	5227279	1073741824	1068667270	0.038364127876090004	0.0527778723133838	141.7428737385299	0

Table A.7: *Bacillus anthracis*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	4214630	4	0	0.9978109584945772	0.9994239346743053	1.3836699632569807E-5	4
2	4214629	16	0	0.9975689437907821	0.9996915354180166	3.283066519699788E-5	16
3	4214628	64	0	0.9952270995210016	0.9996049270129831	8.747539530915498E-5	64
4	4214627	256	0	0.9933102027771378	0.9994171805511821	2.006299552250135E-4	208
5	4214626	1024	0	0.9892298865901743	0.9988662729970227	5.477444136730673E-4	194
6	4214625	4096	0	0.9810751846249667	0.9973012942823112	0.001809901691524341	9
7	4214624	16384	0	0.9641158974086419	0.9926004217296122	0.006649344898262882	0
8	4214623	65536	4	0.930972948232855	0.9780045822855681	0.026690702298926873	0
9	4214622	262144	2917	0.8665389209281402	0.9337572881115214	0.3411947563893245	0
10	4214621	1048576	156167	0.7456347794973736	0.8138128402862368	6.881469022829086	0
11	4214620	4194304	2104873	0.5456321091818479	0.5728321156967433	39.668058111915705	0
12	4214619	16777216	13559007	0.3096657135556026	0.2953320000874121	90.20592987151981	0
13	4214618	67108864	63284052	0.13516290207084014	0.12367738498766455	124.90127631066714	0
14	4214617	268435456	264378689	0.050368752368246006	0.05400608386514907	139.81962963862776	0
15	4214616	1073741824	1069609741	0.01873337926871632	0.030464863575677886	144.92072969758	0

Table A.8: *Bacillus subtilis*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1042519	4	0	0.998339598606836	0.9997004978951441	1.118387410400196E-5	4
2	1042518	16	0	0.9973141950546657	0.9997508574109336	3.203972494435038E-5	16
3	1042517	64	0	0.9943626818555477	0.9993542580889244	1.5114603420931263E-4	56
4	1042516	256	0	0.9904241277831707	0.9987609268128239	4.61443097009407E-4	55
5	1042515	1024	0	0.9821422233732848	0.9970883676939833	0.0016032296282589079	2
6	1042514	4096	0	0.9663381019343625	0.9918321780464182	0.005873836529473902	0
7	1042513	16384	0	0.9332392018133107	0.9750831192871047	0.023617465298467623	0
8	1042512	65536	562	0.8677396519176758	0.9233831825329898	0.3031440458452484	0
9	1042511	262144	34004	0.7437619363248925	0.7872590169811206	6.7573784780181745	0
10	1042510	1048576	510293	0.5366490489299862	0.5311543608362579	42.23577475245041	0
11	1042509	4194304	3376110	0.29394278610544367	0.2555418016891759	94.69139136050151	0
12	1042508	16777216	15816203	0.12179570804252826	0.09388964335677014	128.6404882152732	0
13	1042507	67108864	66095650	0.041678377219529494	0.03070060827579415	142.4744734958034	0
14	1042506	268435456	267405871	0.013006160156392332	0.009675663670203184	147.03469149368286	0
15	1042505	1073741824	1072707329	0.0038676073496050067	0.00296990474546759	148.43073764098298	0

Table A.9: *Chlamydia trachomatis*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	3903260	4	0	0.9887283962636361	0.9984604101509398	3.6772067804541835E-4	4
2	3903258	16	0	0.9884429878834553	0.9984489007219198	6.984339172816571E-4	14
3	3903256	64	0	0.9842516094255668	0.9981928807417899	0.0011027332454463215	44
4	3903254	256	0	0.9811854416853221	0.9979631855422391	0.0016152257323196499	82
5	3903252	1024	0	0.9764391333175516	0.997611490363901	0.002435934029486195	80
6	3903250	4096	0	0.9703477871004932	0.9971422131662896	0.0042313746498980065	29
7	3903248	16384	48	0.9579504043811717	0.9961133099547388	0.012892270848690645	6
8	3903246	65536	2799	0.9336831447467057	0.9933933476262607	0.13831116666043727	0
9	3903244	262144	48670	0.8847917270864952	0.985451759618724	1.3136196282719461	0
10	3903242	1048576	456447	0.7974314685074613	0.9627454817785942	7.139553199291549	0
11	3903240	4194304	2919997	0.6609965054672529	0.9012320557770348	24.388531545799836	0
12	3903238	16777216	14650179	0.48658959561266824	0.7598533937510471	54.97100511386428	0
13	3903236	67108864	64230569	0.311995995117897	0.5297283886438088	89.35129410288694	0
14	3903234	268435456	265058654	0.17542555737114407	0.29938238382965443	116.3997105148667	0
15	3903232	1073741824	1070099203	0.09016937758247523	0.1580468688886179	132.6287101973265	0

Table A.10: *Clostridium botulinum*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	3661391	4	0	0.9985024817070889	0.9999321209161522	6.484188698150766E-6	4
2	3661389	16	0	0.998193308605013	0.9998989013345736	2.311420539793475E-5	16
3	3661387	64	0	0.9963661311956371	0.9998212716940105	6.84845102145242E-5	64
4	3661385	256	0	0.9937261446146745	0.9996282829201882	2.0400438341745694E-4	184
5	3661383	1024	0	0.9889356016565325	0.9992114568774517	6.183363991326879E-4	178
6	3661381	4096	0	0.9808965524210673	0.998170871880864	0.0019971886310934414	10
7	3661379	16384	5	0.9651822441763063	0.9954575598954689	0.007934337565906018	0
8	3661377	65536	778	0.9344066453686687	0.9871995113474157	0.08724003938671519	0
9	3661375	262144	26651	0.8743605885766959	0.9620050445206146	1.1715882254367613	0
10	3661373	1048576	349332	0.766258450040463	0.8937721408586812	8.269999736315084	0
11	3661371	4194304	2653899	0.5960231836653538	0.7324840529209401	32.22747172634221	0
12	3661369	16777216	14293656	0.3874228464817395	0.47180794569087436	73.70631472411281	0
13	3661367	67108864	63975921	0.20344423271417478	0.23495799849668567	112.33413327305436	0
14	3661365	268435456	264995938	0.09040016496579828	0.10297696822188628	133.88454237401197	0
15	3661363	1073741824	1070184731	0.038435413260034634	0.04922061482087727	142.78612907760254	0

Table A.11: *Desulfovibrio vulgaris*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	4639675	4	0	0.999161794737778	0.9936847914266551	2.2578765528629547E-6	4
2	4639674	16	0	0.9980668900444298	0.9994935232871515	1.8906555363975173E-5	16
3	4639673	64	0	0.9964788466773413	0.9994431030121095	5.3677747222001757E-5	64
4	4639672	256	0	0.9946487596536997	0.9991937932266025	1.4569948601348155E-4	233
5	4639671	1024	0	0.9904952312351458	0.9985524826901533	4.3223885790498554E-4	282
6	4639670	4096	0	0.9826211777992832	0.9966490857124178	0.0014898292288835929	0
7	4639669	16384	1	0.9666801661928901	0.9912336974929288	0.00577045319200499	0
8	4639668	65536	176	0.9360378802966074	0.9750925692667044	0.030855734572379608	0
9	4639667	262144	5617	0.8753511836086512	0.9279010285209854	0.29926024554012026	0
10	4639666	1048576	150468	0.7608580014164813	0.8074051847835827	5.4340487585481805	0
11	4639665	4194304	1997469	0.5666611705802036	0.5839611672764892	35.469796867214335	0
12	4639664	16777216	13298293	0.3293018632383724	0.34189694488904626	86.84660197742494	0
13	4639663	67108864	62938541	0.14963328155514743	0.19032694717147483	122.98516663464349	0
14	4639662	268435456	264003264	0.06243515152612411	0.12216061217328313	138.28518591510127	0
15	4639661	1073741824	1069224203	0.030120735114052555	0.0931563648723278	143.4853314995053	0

Table A.12: *Escherichia coli*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1830023	4	0	0.9963557835065461	0.9995271232137565	4.0333533406255114E-5	4
2	1830022	16	0	0.9959678080372805	0.999643608247859	8.006216009294775E-5	16
3	1830021	64	0	0.9936618213670773	0.9995849931737157	1.6551515852819223E-4	60
4	1830020	256	0	0.9917984502901608	0.9994023525333484	3.3786218213548836E-4	127
5	1830019	1024	0	0.9859067037008905	0.998791010111913	9.875109447393341E-4	44
6	1830018	4096	0	0.9746904128811847	0.9970026467898533	0.0033916997690324503	0
7	1830017	16384	12	0.951560559273493	0.9912005697421815	0.014397978582774668	0
8	1830016	65536	1077	0.9058631181366721	0.9737432813781214	0.17573481456962653	0
9	1830015	262144	33867	0.8196708770146692	0.9263723616480362	2.9186693069340435	0
10	1830014	1048576	442423	0.6703489700078797	0.814232471065549	19.721358446704784	0
11	1830013	4194304	3083246	0.46206666291441645	0.6107052736466104	59.43368161820964	0
12	1830012	16777216	15280705	0.2587212542868571	0.38195499407947764	102.51188182927962	0
13	1830011	67108864	65423708	0.1260560728869936	0.22394208995671536	128.18686590096814	0
14	1830010	268435456	266679228	0.06495210408686292	0.1437434796679517	138.7164353114016	0
15	1830009	1073741824	1071960397	0.04194405601283924	0.10665114556380303	142.4188779580965	0

Table A.13: *Haemophilus influenzae*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1667825	4	0	0.9913150360499453	0.9968385353637794	2.1778176371570654E-4	4
2	1667824	16	0	0.9904552278897534	0.9980413619133174	4.6415614474230833E-4	12
3	1667823	64	0	0.9879249776505061	0.9978171708428498	8.566035483578659E-4	50
4	1667822	256	0	0.9846404472419719	0.9976870719722948	0.001341962859520012	80
5	1667821	1024	0	0.9787345284655847	0.9972444863782457	0.0023687059383820518	24
6	1667820	4096	2	0.9677567123550503	0.9959233827728938	0.005761147576515556	9
7	1667819	16384	192	0.9458184611159844	0.9920194790960117	0.03402003715210895	0
8	1667818	65536	4290	0.9031848798849754	0.9808388513054596	0.3363406515867059	0
9	1667817	262144	58657	0.8216740805496047	0.9492448844692574	3.413961884147968	0
10	1667816	1048576	538038	0.6848345381025245	0.8654211578695782	18.8218378949124	0
11	1667815	4194304	3275984	0.49689803725233317	0.6837120168074035	52.37240871268567	0
12	1667814	16777216	15509148	0.30872027696133986	0.4277532143727841	90.94628936372527	0
13	1667813	67108864	65628847	0.1684133652873554	0.21747109729312877	119.28403890905687	0
14	1667812	268435456	266855923	0.08912335443083508	0.10826941071030896	134.23140048696868	0
15	1667811	1073741824	1072122187	0.052487961765451896	0.064129364178643	140.63269316759192	0

Table A.14: *Helicobacter pylori*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	580076	4	0	0.9903047186920334	0.9978764406804925	3.101552386611061E-4	4
2	580075	16	0	0.9885566521570487	0.9980146244018896	6.564448213954027E-4	14
3	580074	64	0	0.9797232766853884	0.9971429365389813	0.0016289793137196154	12
4	580073	256	0	0.974220486042274	0.9959560428424044	0.003253581630718489	10
5	580072	1024	0	0.962877022162766	0.9941995437883543	0.006933964400253681	0
6	580071	4096	14	0.9442792347833282	0.9905194050457541	0.026182953982611543	0
7	580070	16384	851	0.9065147309807438	0.9807230507428708	0.2915835112519192	0
8	580069	65536	14196	0.8361143243303814	0.9530126319657017	2.6684560747631245	0
9	580068	262144	126707	0.7173296923808933	0.8794093174396036	13.632568169150076	0
10	580067	1048576	779814	0.5436044456933423	0.7135059784298358	41.696826976403855	0
11	580066	4194304	3789892	0.3449607458461623	0.44984943328729543	82.07096278943135	0
12	580065	16777216	16281022	0.17889374466654595	0.21194701575783104	116.65262530884286	0
13	580064	67108864	66568079	0.07742938710211289	0.08252574038510997	136.07524451293065	0
14	580063	268435456	267876985	0.029231307633825976	0.02992271724885458	144.40698627283317	0
15	580062	1073741824	1073176778	0.010057545572714655	0.010864596043172126	147.48935826614883	0

Table A.15: *Mycoplasma genitalium*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	6264404	4	0	0.991501505969283	0.998518056510378	2.0852801516339211E-4	4
2	6264403	16	0	0.9909004896396353	0.9986249183583796	4.672575937778555E-4	16
3	6264402	64	0	0.9853623697840592	0.9984222205487662	8.79018562087924E-4	46
4	6264401	256	0	0.9825755088156074	0.998056430961034	0.0014014661073638375	99
5	6264400	1024	0	0.9778545431326224	0.9977093507447125	0.0021129914480761446	144
6	6264399	4096	0	0.9730931889874831	0.9971666964810718	0.003428285042409439	78
7	6264398	16384	0	0.9634135634421696	0.9960594291551214	0.0069619102714149235	12
8	6264397	65536	276	0.9442746045628972	0.9935415649847638	0.036284492659049414	0
9	6264396	262144	21387	0.9047668761681095	0.9865672533444223	0.6663180396205012	0
10	6264395	1048576	326432	0.8312207324091153	0.9683554712959311	4.6889579420434755	0
11	6264394	4194304	2536746	0.7104968174096329	0.918255364674384	17.74333002512336	0
12	6264393	16777216	13827364	0.5483790688100187	0.7979236513648378	43.36840448407536	0
13	6264392	67108864	62860550	0.371762175802536	0.5897959326547151	77.24494573159515	0
14	6264391	268435456	263226896	0.2200802919230297	0.3404055704669516	107.89661511066056	0
15	6264390	1073741824	1067974626	0.11482809978305952	0.1615690546694321	128.60620422662012	0

Table A.16: *Pseudomonas aeruginosa*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1908256	4	0	0.9995881055791257	0.9999824585641273	1.0561481960331381E-6	4
2	1908255	16	0	0.9986699890737873	0.999941269389624	8.382618148705937E-6	16
3	1908254	64	0	0.9969417069216152	0.9998288452416673	4.211714737768137E-5	62
4	1908253	256	0	0.9937041891195769	0.9995009687216164	1.9986813656286417E-4	153
5	1908252	1024	0	0.9870006686747872	0.9984892085674628	8.333554724255644E-4	30
6	1908251	4096	0	0.9748675619716693	0.9955906804487865	0.0033419862887312615	0
7	1908250	16384	5	0.9512212760382549	0.98705328338288	0.014486003427482443	0
8	1908249	65536	777	0.9049741412153236	0.96075496083238	0.15897321885244725	0
9	1908248	262144	26983	0.8157493156025841	0.8851199806520874	2.5172093573175474	0
10	1908247	1048576	393714	0.6586238573937231	0.7139738967605394	19.610758981613586	0
11	1908246	4194304	2976230	0.43458547797296576	0.4617418156268787	64.15029121678653	0
12	1908245	16777216	15162064	0.22148675877573376	0.25312816096308005	109.35889069810334	0
13	1908244	67108864	65317621	0.09369661322137002	0.14684511768244207	133.25669088662895	0
14	1908243	268435456	266582306	0.03986389574074156	0.10170839709570216	142.27415814485752	0
15	1908242	1073741824	1071868431	0.02068919979750994	0.08183396108771063	145.3015093804829	0

Table A.17: *Pyrococcus furiosus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2813862	4	0	0.9976544691957175	0.9998326238967088	4.34087487788978E-5	4
2	2813860	16	0	0.9958277952705537	0.999758596576658	1.0903532029620053E-4	16
3	2813858	64	0	0.9926542135388495	0.999574299416066	2.324492898140746E-4	56
4	2813856	256	0	0.9902347525957262	0.9993677239955902	4.457360460438974E-4	119
5	2813854	1024	0	0.9856680552722351	0.9989644864183236	9.934614892493164E-4	80
6	2813852	4096	0	0.9766316778565468	0.9979019516031594	0.0028961911541932664	16
7	2813850	16384	2	0.9589814666737744	0.9950162174936944	0.01007731776307634	0
8	2813848	65536	797	0.9241426686871501	0.986281450827879	0.10772732054330528	0
9	2813846	262144	29444	0.8559715066140791	0.9593690995590278	1.6828146973752922	0
10	2813844	1048576	389870	0.7354601036873402	0.8851628094223768	11.823224628240755	0
11	2813842	4194304	2849198	0.5537205002981689	0.7124250653828352	40.96215649927552	0
12	2813840	16777216	14753569	0.34626098143462314	0.44961675916174426	83.10908353464261	0
13	2813838	67108864	64653361	0.17936640275666194	0.23104164639795416	117.30822123668234	0
14	2813836	268435456	265783231	0.08314023987183328	0.1197679671395349	135.29279645974313	0
15	2813834	1073741824	1071016349	0.04080695591850836	0.07622512915201868	142.47057013610743	0

Table A.18: *Staphylococcus aureus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2030921	4	0	0.9949978359571839	0.9992795075220364	7.761314058917908E-5	4
2	2030920	16	0	0.9943316329545231	0.9993278465542141	1.6801731300384142E-4	16
3	2030919	64	0	0.9917392077182793	0.9990645975760012	3.382629546585828E-4	58
4	2030918	256	0	0.9885155382935205	0.9987487482698395	5.92544835955388E-4	90
5	2030917	1024	0	0.9833045860564464	0.9981000559307539	0.0012605761057521742	42
6	2030916	4096	0	0.9730643709537962	0.9962803557411681	0.003597411962480166	0
7	2030915	16384	0	0.9515454856554804	0.9910194096379222	0.01240157991691608	0
8	2030914	65536	417	0.9093137375585574	0.9751367569057182	0.11595271726554271	0
9	2030913	262144	26027	0.8255321621359457	0.9264853387708629	2.4473541643964745	0
10	2030912	1048576	404415	0.6782814814231242	0.799624577091774	18.37220102994894	0
11	2030911	4194304	2990782	0.46772507510176464	0.5579424301710008	57.539144823236214	0
12	2030910	16777216	15131943	0.2592566878886804	0.296442708226934	101.2974577435672	0
13	2030909	67108864	65239790	0.11975130348036278	0.14066150572810998	128.48245566075056	0
14	2030908	268435456	266480177	0.05346180132236422	0.0770881046885005	140.04990346266467	0
15	2030907	1073741824	1071757724	0.02855177514283025	0.055327597068715824	144.04999096241164	0

Table A.19: *Streptococcus mutans*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2221315	4	0	0.9975865647150449	0.9997162857974253	2.097434045774643E-5	4
2	2221314	16	0	0.9958920710894542	0.9994686471595192	8.974996022465917E-5	16
3	2221313	64	0	0.9929118498833798	0.9992041471804838	2.2519241077935458E-4	58
4	2221312	256	0	0.9890852793304138	0.9985888326116023	6.006661794232606E-4	114
5	2221311	1024	0	0.9825539962661689	0.9970474855479906	0.001823412307100714	44
6	2221310	4096	0	0.9713241285547717	0.9934492439838425	0.005128488475366213	0
7	2221309	16384	0	0.9486235368424654	0.9831994319798478	0.01625839636015805	0
8	2221308	65536	391	0.9059000372753351	0.9543789676988104	0.10104805614315214	0
9	2221307	262144	18807	0.8216414930489122	0.8791267558729262	1.805223488494659	0
10	2221306	1048576	343650	0.6715418767157699	0.7152720838615902	17.036827222413315	0
11	2221305	4194304	2846941	0.45428430584723845	0.49065814592524554	59.44044779605143	0
12	2221304	16777216	14960497	0.24430064502652493	0.31546838809052075	104.02483715964703	0
13	2221303	67108864	65078889	0.11599047946182939	0.22497728173195738	128.36450985946487	0
14	2221302	268435456	266329475	0.06046859004313687	0.18492088821874658	137.7662183080229	0
15	2221301	1073741824	1071610432	0.04084723322053152	0.16557163683345227	140.89696120235257	0

Table A.20: *Streptococcus pneumoniae*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	949626	4	0	0.9967144960226447	0.9997755900124569	3.1180508116367E-5	4
2	949625	16	0	0.9955645649598526	0.9996497203743084	9.215851127576698E-5	16
3	949624	64	0	0.9926223431589766	0.9994761295388405	2.2198175388779762E-4	48
4	949623	256	0	0.9894326485352608	0.9992157879552837	5.404493131086606E-4	35
5	949622	1024	0	0.9812009410060003	0.9983783638024313	0.0017354493262728577	6
6	949621	4096	0	0.9655462547690078	0.9959375201542778	0.006382804181886499	0
7	949620	16384	45	0.9348307744150292	0.9890136083382328	0.04528338081682246	0
8	949619	65536	3767	0.8753120988522766	0.9700924084565002	0.8487172077714453	0
9	949618	262144	72322	0.7641072515474643	0.9229849943285295	8.275467388143607	0
10	949617	1048576	631644	0.5888774105771064	0.8236042077475467	34.49545136388268	0
11	949616	4194304	3537900	0.3794691749085946 6	0.6611757528680389	76.41688136604759	0
12	949615	16777216	15962553	0.20538218119974938	0.4751948998421617	112.32512637661114	0
13	949614	67108864	66220209	0.10265855389663592	0.31977853637317266	131.75822326146368	0
14	949613	268435456	267517172	0.05497607972932128	0.21429933879956273	140.01037729443476	0
15	949612	1073741824	1072811872	0.0350416801809581	0.14595216822862198	143.2935670100155	0

Table A.21: *Candida albicans*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	38047924	4	0	0.9995165045010077	0.897823620621571	1.013066596931652E-6	4
2	38047100	16	0	0.9993783757500572	0.9997392761407228	2.3825495882062084E-6	16
3	38046275	64	0	0.9987931275795068	0.9997724570437846	5.689543966932842E-6	64
4	38045450	256	0	0.9982857871309184	0.9997652525876801	1.4198527581163293E-5	256
5	38044627	1024	0	0.9967489495954317	0.9995648191612441	4.798789484897236E-5	1024
6	38043804	4096	0	0.9940042010520294	0.999072511063084	1.6893093384801744E-4	3160
7	38042982	16384	0	0.9882902975376641	0.997889191758642	6.440072700523661E-4	484
8	38042159	65536	0	0.9769887403078253	0.9948870074790318	0.0025536104726960056	14
9	38041336	262144	0	0.9544035467103469	0.9881899811129488	0.010121100197988524	2
10	38040511	1048576	6	0.9100043635060528	0.976010798224927	0.041064537729834406	2
11	38039687	4194304	38980	0.8229558250571305	0.9577857130325643	0.6626849686723125	2
12	38038864	16777216	3427164	0.660532764595704	0.9363237088047023	17.405467318720365	2
13	38038042	67108864	41759894	0.41822178964942514	0.9142699325399566	69.40442961652752	2
14	38037220	268435456	235829289	0.20463814127320556	0.8917206266681069	113.16162560048036	2
15	38036398	1073741824	1038168288	0.0928696770919265	0.8681828082911317	133.11177233355346	0

Table A.22: *Neurospora crassa*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	231027417	4	0	0.9995031542078835	0.9999937415530492	1.6373639986924003E-6	4
2	231027400	16	0	0.9992182528998724	0.9999943446838481	3.614091507579643E-6	16
3	231027383	64	0	0.9988615765084435	0.9999942588077917	5.737462131545716E-6	64
4	231027366	256	0	0.9986591285467021	0.9999933239192526	8.852127860787978E-6	256
5	231027349	1024	0	0.9980535767650608	0.9999903782558304	1.7368798953200707E-5	1024
6	231027332	4096	0	0.9970288277406069	0.9999821995113923	4.6124034570082224E-5	4016
7	231027315	16384	0	0.9948760041642695	0.9999637304319334	1.531189881092605E-4	9492
8	231027298	65536	0	0.9906828802542633	0.9999184257119983	5.573064851521707E-4	8537
9	231027281	262144	0	0.9824282873328714	0.9998135187101063	0.002102026726602998	4822
10	231027264	1048576	0	0.9667298055349866	0.9996037213955359	0.007911540486701448	2033
11	231027247	4194304	8469	0.9364346794990809	0.9992464294474778	0.048224144664767814	1236
12	231027230	16777216	1145385	0.8788923842440565	0.9987290933707477	1.2583122496978507	296
13	231027213	67108864	22973502	0.7765802723854873	0.998080748806917	10.644126979748455	132
14	231027196	268435456	182074571	0.6274407494431955	0.9973163097797467	33.86075435699718	82
15	231027179	1073741824	945564276	0.46445215002170803	0.9964398307562391	63.260998720719606	60

Table A.23: *Apis mellifera*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	127486206	4	0	0.999792918772718	0.9999959706207288	1.2932357856135322E-7	4
2	127486129	16	0	0.9997184634886828	0.9999962029668423	3.888622036446186E-7	16
3	127486052	64	0	0.9993976125325459	0.9999928694050713	1.5789274498530528E-6	64
4	127485975	256	0	0.9988710287543394	0.9999800618905736	6.322409932794663E-6	256
5	127485898	1024	0	0.9979426587244967	0.99995501082963	2.2057847214558456E-5	1024
6	127485821	4096	0	0.9963079109793708	0.9998908054074234	7.578498323955266E-5	4024
7	127485744	16384	0	0.9930597102684674	0.9997137118669932	2.709622574831945E-4	8610
8	127485667	65536	0	0.9867627315312238	0.9992407735252086	0.0010051786615203754	2649
9	127485590	262144	0	0.9742605419169336	0.9979840906022217	0.003747531521874544	362
10	127485513	1048576	0	0.9501957606743913	0.9949524450676315	0.014102870425750924	60
11	127485436	4194304	3230	0.9036333060036756	0.9883809597121085	0.06937488199225342	8
12	127485359	16777216	886701	0.8151168715773864	0.9760082948739133	2.150640725000231	6
13	127485282	67108864	23200293	0.6587269893633683	0.9558110821129785	22.035403043898235	4
14	127485205	268435456	191655240	0.4481081471375443	0.9265101713039868	64.29590481560717	4
15	127485128	1073741824	974937551	0.26542368141953	0.8877858166686977	100.96362686156921	4

Table A.24: *Drosophila melanogaster*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2466956460	4	0	0.9992052336424292	0.9999394204984507	2.098934235454l466E-6	4
2	2466956430	16	0	0.9991111444153069	0.9999640667991094	4.1515158915842335E-6	16
3	2466956400	64	0	0.9988175786163063	0.9999690790675032	6.352346410991031E-6	64
4	2466956370	256	0	0.9986224085511493	0.9999705571444946	8.764048743308E-6	256
5	2466956340	1024	0	0.9984038225824459	0.9999705683377873	1.1537742779403444E-5	1024
6	2466956310	4096	0	0.998131562775832	0.9999693547722385	1.594976051421794E-5	4092
7	2466956280	16384	0	0.9976415552852845	0.999964992210905	2.736980892257685E-5	15942
8	2466956250	65536	0	0.9965791355237856	0.9999532622230635	6.64616296297994E-5	50711
9	2466956220	262144	0	0.9943211639159125	0.9999277796964797	2.105122932884605E-4	125380
10	2466956190	1048576	0	0.9897810799793733	0.9998874064061525	7.521298292529183E-4	94938
11	2466956160	4194304	843	0.9808823177465789	0.9998396683634909	0.002969291319677912	27372
12	2466956130	16777216	162015	0.9636897235785056	0.9997897683386373	0.027396134356592147	16745
13	2466956100	67108864	4955837	0.9307904717072184	0.9997396677301307	0.3418841487107447	14552
14	2466956070	268435456	69380851	0.8698238351686578	0.9996886799525687	2.643729637119704	13549
15	2466956040	1073741824	553034323	0.7660147839521292	0.9996341495196697	11.674945743510245	12980

Table A.25: *Bos taurus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2309875278	4	0	0.9997974773770447	0.999996157809793	1.5334935525791478E-7	4
2	2309875239	16	0	0.9997422068560474	0.9999974762638534	2.8540864521826347E-7	16
3	2309875200	64	0	0.9996990616635911	0.999997661411679	4.4530442203779E-7	64
4	2309875161	256	0	0.9996046318799304	0.9999973781287766	7.798648814934986E-7	256
5	2309875122	1024	0	0.9994475865868908	0.999996683859761	1.5177835567443026E-6	1024
6	2309875083	4096	0	0.9991268623074714	0.9999947416202218	3.873744301710922E-6	4096
7	2309875044	16384	0	0.9984876385373858	0.9999898640749422	1.2441811333200383E-5	16070
8	2309875005	65536	0	0.9972062150609747	0.9999786034445141	4.5402212777384616E-5	51151
9	2309874966	262144	0	0.9946816125630932	0.9999573043372154	1.7335995175683956E-4	128658
10	2309874927	1048576	0	0.9897626656215919	0.9999235690021896	6.788812539803044E-4	90860
11	2309874888	4194304	472	0.9801872515962865	0.9998792222987728	0.0028046680331142663	19776
12	2309874848	16777216	120750	0.9616833158399758	0.999826643056101	0.025425958336724275	8434
13	2309874809	67108864	4312343	0.9260025736745459	0.9997651598937968	0.3400285035054528	5962
14	2309874770	268435456	65014719	0.8592017038222379	0.9996919015630809	2.818682026734249	4945
15	2309874731	1073741824	534169161	0.7438189582065262	0.9996043994436747	12.521587059641304	4316

Table A.26: *Canis familiaris*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2335454483	4	0	0.9997218669836093	0.9999947605827264	2.2333594949330305E-7	4
2	2335454451	16	0	0.9996829880370037	0.9999960693630487	4.819812387113075E-7	16
3	2335454419	64	0	0.9995895192848977	0.9999963221884891	7.709511043802197E-7	64
4	2335454387	256	0	0.9994899258120257	0.9999961290309034	1.1935202108107767E-6	256
5	2335454355	1024	0	0.9993335913430857	0.9999953691177975	2.0336272194570586E-6	1024
6	2335454323	4096	0	0.9990435347084285	0.9999926559738093	4.49642955356383E-6	4096
7	2335454291	16384	0	0.998405504653056	0.999985534421487	1.3135939846837636E-5	16172
8	2335454259	65536	0	0.9971504592845892	0.9999657392890334	4.5299405220132386E-5	54182
9	2335454227	262144	0	0.9945898305974378	0.9999092482138143	1.7076620043637075E-4	131614
10	2335454195	1048576	0	0.9895867296168487	0.9997635952308795	6.665650070050442E-4	99825
11	2335454163	4194304	245	0.9797653939226552	0.9994400623554867	0.0027112069376402604	18542
12	2335454131	16777216	76778	0.9606667787730574	0.9988655924505822	0.02061123032558792	5301
13	2335454099	67108864	3368226	0.9237715958210317	0.9980544313996045	0.28681074144087193	3056
14	2335454067	268435456	55840802	0.854312877393876	0.9970708662036708	2.6620973947159245	2276
15	2335454035	1073741824	503802794	0.733514307850636	0.9959763338853489	13.245520392205773	1796

Table A.27: *Equus caballus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2870607502	4	0	0.9994600898942401	0.9999769899134303	9.60279956629913E-7	4
2	2870607478	16	0	0.9994023097852461	0.9999844801856764	1.8236249509101997E-6	16
3	2870607454	64	0	0.9992228606537994	0.9999862174846489	2.7522106628680826E-6	64
4	2870607430	256	0	0.9990708753930871	0.9999860555366059	3.899617318160077E-6	256
5	2870607406	1024	0	0.9988853014197233	0.9999851780988541	5.55886195429443E-6	1024
6	2870607382	4096	0	0.9986125702786895	0.9999830020926509	9.072841840359257E-6	4094
7	2870607358	16384	0	0.9980637351937004	0.9999778133939385	1.94026243725485E-5	15976
8	2870607334	65536	0	0.9969895140663638	0.9999664791276693	5.3243721382887844E-5	52114
9	2870607310	262144	0	0.994779575754651	0.9999466347703247	1.7718906176448902E-4	136718
10	2870607286	1048576	0	0.9904267483281236	0.9999189820612984	6.427783511111519E-4	134194
11	2870607262	4194304	979	0.9819811826282491	0.9998868220314195	0.0025767159905912868	36148
12	2870607238	16777216	167126	0.9657469479981852	0.9998510044725942	0.023864649920525654	16900
13	2870607214	67108864	4801682	0.9347942696280007	0.9998091994143211	0.2702850977845962	12802
14	2870607190	268435456	65721695	0.877325748633689	0.9997585213899126	2.1700965549552302	11349
15	2870607166	1073741824	527172585	0.7781341921167628	0.9996959554419126	9.485419708235383	10244

Table A.28: *Homo sapiens*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2617133082	4	0	0.9996618200250926	0.9999883276129492	4.05298452151317E-7	4
2	2617133061	16	0	0.9995950320540465	0.9999931001124631	7.975099233791554E-7	16
3	2617133040	64	0	0.9994750370046148	0.9999939610896901	1.1996871302654513E-6	64
4	2617133019	256	0	0.9993964361044959	0.9999939939367889	1.7354486748076928E-6	256
5	2617132998	1024	0	0.9992172266363362	0.999993304696564	2.669672261037786E-6	1024
6	2617132977	4096	0	0.9989494695056911	0.9999909225171774	5.2881150454378E-6	4090
7	2617132956	16384	0	0.9983525498809239	0.99998484408565	1.4150117194119353E-5	15932
8	2617132935	65536	0	0.9971190725930779	0.9999718288352973	4.7044126530524474E-5	50875
9	2617132913	262144	0	0.9946386945308344	0.9999499599865747	1.7467485761141622E-4	134416
10	2617132892	1048576	1	0.9898410645935208	0.9999236267835026	6.726803139351897E-4	115262
11	2617132871	4194304	1481	0.9806322087954853	0.9998987701360654	0.00284679210880223	28176
12	2617132850	16777216	193368	0.9630770715365099	0.9998763742526109	0.028204097606792773	16785
13	2617132829	67108864	5175726	0.9297968429557306	0.9998563701159381	0.3104934008107785	14188
14	2617132808	268435456	68112990	0.8684692022706094	0.9998366660152772	2.435000011399865	12890
15	2617132787	1073741824	536384356	0.7635794255173189	0.9998169022903823	10.53352077127197	12038

Table A.29: *M musculus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2725181837	4	0	0.9997916729840586	0.999996939827133	1.3299244824002578E-7	4
2	2725181816	16	0	0.9997677413681965	0.9999977255583663	2.6275686425946404E-7	16
3	2725181795	64	0	0.9996858451786333	0.9999976022471317	5.23982469224268E-7	64
4	2725181774	256	0	0.9995228901747382	0.9999965696428046	1.1806553595240116E-6	256
5	2725181753	1024	0	0.9991890548226491	0.9999936171485589	3.327491677178407E-6	1024
6	2725181732	4096	0	0.9987309283049312	0.9999864079213738	9.794041894971152E-6	4096
7	2725181711	16384	0	0.998013341650523	0.999970731396838	3.0139406156919902E-5	15964
8	2725181690	65536	0	0.9967242723548462	0.9999422614632844	9.289152195536131E-5	51499
9	2725181669	262144	0	0.9943085243907092	0.9999020066922998	2.8111181715428105E-4	133454
10	2725181648	1048576	0	0.9897451301932443	0.9998589357250509	8.507823276855586E-4	119998
11	2725181627	4194304	908	0.980997048972105	0.9998189592691157	0.0029202956929848625	29602
12	2725181606	16777216	155910	0.9642496038482362	0.9997803313865362	0.024611182804885837	12220
13	2725181585	67108864	4653759	0.9322203151464492	0.9997389932231178	0.28257176489627733	8880
14	2725181564	268435456	64729101	0.8725020488212872	0.9996899477117211	2.307956692767757	7731
15	2725181543	1073741824	526098599	0.7691490658242727	0.9996295327791973	10.278030899391615	6952

Table A.30: *Macaca mulatta*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	3412593428	4	0	0.9998745177211893	0.03698370404918901	5.027681779776772E-8	4
2	3412593419	16	0	0.999832914171133	0.09426879170661247	1.0342088255299583E-7	16
3	3412593410	64	0	0.9997909097527091	0.1562017132846937	2.019278700516451E-7	64
4	3412593401	256	0	0.9997165355240631	0.21766865777204833	3.9675584669318697E-7	256
5	3412593392	1024	0	0.9995805729439213	0.28174822918799136	8.835763940596765E-7	1024
6	3412593383	4096	0	0.9993340024008363	0.34767973144128556	2.4100368554337607E-6	4092
7	3412593374	16384	0	0.9988141968419587	0.41956576314807953	8.207740925135233E-6	15696
8	3412593365	65536	0	0.9978153175012107	0.5008823318231559	3.061124731998886E-5	47001
9	3412593356	262144	0	0.9958471559539659	0.6027133107778588	1.1891444282433878E-4	140322
10	3412593347	1048576	1	0.9920573955804527	0.7286533475791662	4.686318252000874E-4	189154
11	3412593338	4194304	4006	0.9847775726977054	0.8561469868779267	0.0023341928599566394	72328
12	3412593329	16777216	374998	0.9708762661066551	0.941749358981254	0.028846182387473506	28861
13	3412593320	67108864	7479876	0.9445949176270438	0.9804824438525646	0.29465011053405915	18420
14	3412593311	268435456	83702859	0.8962853859968783	0.9938635414112688	1.8269089097517697	14957
15	3412593302	1073741824	571949809	0.8130116209200717	0.9979330297021037	6.601222295229983	13028

Table A.31: *Monodelphis domestica*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	409491489	4	0	0.9994708681234642	0.9999669633404319	8.163915718756335E-7	4
2	409491470	16	0	0.9994029863430366	0.9999826521122326	1.6207876683178696E-6	16
3	409491451	64	0	0.999248152802096	0.9999851895729501	2.5702520722431476E-6	64
4	409491432	256	0	0.9990494599652576	0.9999845684817582	4.092956543031127E-6	256
5	409491413	1024	0	0.9986901190526308	0.9999808443564991	7.72416381089951E-6	1024
6	409491393	4096	0	0.998015401510527	0.999969865436245	1.9769815473234895E-5	4047
7	409491374	16384	0	0.9964493611042463	0.9999399254078716	6.746099389856098E-5	13046
8	409491355	65536	0	0.9933765927732467	0.9998800233177287	2.54538652605292E-4	25885
9	409491336	262144	0	0.9873213923139024	0.9997838097530879	9.930468937576317E-4	9168
10	409491317	1048576	14	0.9755750816078964	0.9996568237912223	0.003962178335298152	2947
11	409491298	4194304	13701	0.9529130653223308	0.9995053646034257	0.028009333908990946	2150
12	409491279	16777216	897097	0.9095627186726973	0.9993261178071648	0.508680138387665	1783
13	409491260	67108864	16611062	0.8295249378460483	0.9991113843164686	4.528219996382577	1514
14	409491241	268435456	145826825	0.6978826831609812	0.9988506795193083	20.84019512631868	1376
15	409491222	1073741824	864924945	0.5259547370712625	0.9985422207349702	54.972922546002536	1226

Table A.32: *Ornithorhynchus anatinus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2858821973	4	0	0.9995069497109955	0.9999800192060283	8.499492843648238E-7	4
2	2858821947	16	0	0.999442089073902	0.9999857692521031	1.6157149414871487E-6	16
3	2858821921	64	0	0.9992802682164686	0.9999866462673787	2.464874186747614E-6	64
4	2858821895	256	0	0.9991325549155975	0.9999860057876637	3.568156687571977E-6	256
5	2858821869	1024	0	0.9989416454964162	0.999984407206211	5.263509269895534E-6	1024
6	2858821843	4096	0	0.9986139923305463	0.999980658378099	9.070377428985214E-6	4096
7	2858821817	16384	0	0.9980100428203778	0.9999721073957103	2.0279503986983694E-5	15970
8	2858821791	65536	0	0.996870793405814	0.999953921652155	5.6844268618023055E-5	51982
9	2858821765	262144	0	0.994598104999386	0.9999241140992666	1.8580483554439593E-4	136372
10	2858821739	1048576	0	0.9901667478540186	0.9998855166587303	6.593477776229542E-4	132452
11	2858821713	4194304	868	0.9815620069064447	0.9998437858731283	0.0025994297197623095	33672
12	2858821687	16777216	157792	0.9650595371322996	0.9998004526424245	0.023767905420033625	14264
13	2858821661	67108864	4736008	0.9335704624073785	0.9997512509488299	0.2754044248569349	10256
14	2858821635	268435456	65494935	0.8750414647677032	0.9996913109634592	2.216898584690755	8771
15	2858821609	1073741824	527202307	0.7740416817312507	0.9996178254112592	9.693063525193114	7828

Table A.33: *Pan troglodytes*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	2504457701	4	0	0.999926082600666	0.9999994984812131	1.70081282896401E-8	4
2	2504457680	16	0	0.99988227950412	0.9999994459730416	7.35194015650942E-8	16
3	2504457659	64	0	0.9998187851176604	0.9999992997759637	1.6851048899977661E-7	64
4	2504457638	256	0	0.9997208185958544	0.9999988812428817	3.9697379889228473E-7	256
5	2504457617	1024	0	0.9995236920793138	0.9999977495974366	1.1142214704555299E-6	1024
6	2504457596	4096	0	0.9991767958046913	0.9999947743488631	3.5697581703710176E-6	4096
7	2504457575	16384	0	0.998502935311252	0.9999878066747736	1.2420817371256455E-5	16024
8	2504457554	65536	0	0.9972160570304479	0.9999735094967741	4.5664072741944035E-5	52193
9	2504457533	262144	0	0.9946962738137992	0.9999523732168353	1.7625534218061162E-4	132286
10	2504457512	1048576	1	0.9898095588055638	0.999930189245081	6.91599865983315E-4	102627
11	2504457491	4194304	706	0.9804001173202584	0.9999122755916817	0.0028403654856564253	27702
12	2504457470	16777216	127985	0.9623728906843845	0.9998977583962455	0.024512257487243415	18732
13	2504457449	67108864	4288671	0.9279634209508983	0.9998854245708302	0.2995677039608326	16562
14	2504457428	268435456	62104766	0.8640358018495333	0.9998736427162348	2.5346823818136732	15745
15	2504457407	1073741824	521923243	0.7540457357037416	0.9998619883102272	11.755066740619364	15184

Table A.34: *Rattus norvegicus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	89693294	4	0	0.9997734055792399	0.9999975762454479	4.153994050431895E-7	4
2	89693288	16	0	0.9995990000946336	0.9999964704648688	2.2229125857015666E-6	16
3	89693282	64	0	0.9990969223313737	0.9999949929927727	5.999669807127362E-6	64
4	89693276	256	0	0.998595758727778	0.9999913708978508	1.3824250932803658E-5	256
5	89693270	1024	0	0.9972764957727598	0.9999771344920334	4.40848419044804E-5	1012
6	89693264	4096	0	0.9950464507568818	0.999938591987997	1.4927272885897533E-4	3363
7	89693258	16384	0	0.9908747433391258	0.9998285592768176	5.310710253044977E-4	4002
8	89693252	65536	0	0.9834292439301899	0.9995342706171546	0.0018272843208978241	1911
9	89693246	262144	0	0.9694384792362181	0.9986937907890053	0.006274341261677808	676
10	89693240	1048576	72	0.9431992199189148	0.9964016202379163	0.022492903837597274	161
11	89693234	4194304	72932	0.8936048175049637	0.9905619525447471	0.35603462771201666	30
12	89693228	16777216	3253715	0.8025430749353786	0.9773587881044947	5.805999421425932	13
13	89693222	67108864	36628925	0.6568737602045336	0.9518470689749292	26.727999867097736	2
14	89693216	268435456	219629128	0.48210854653712054	0.9111948156075493	58.922192647020076	2
15	89693210	1073741824	1010903999	0.3284786886320603	0.8587895548737233	88.25052414397467	0

Table A.35: *Caenorhabditis elegans*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	989168363	4	0	0.9996097418655514	0.9999850804403569	5.617152330858426E-7	4
2	989168330	16	0	0.9995227890080144	0.999990751163669	1.0913742098446305E-6	16
3	989168297	64	0	0.9993715134200263	0.9999912930096884	1.7004563278625215E-6	64
4	989168264	256	0	0.9992549528459195	0.9999902239583823	2.7261872672553177E-6	256
5	989168230	1024	0	0.9990042320708177	0.9999880190499499	4.627457279409672E-6	1024
6	989168197	4096	0	0.9985605966666556	0.999982086023565	1.023354536780052E-5	4074
7	989168164	16384	0	0.9975952400344337	0.9999651589239266	3.052724798096925 6E-5	14964
8	989168131	65536	0	0.9956112577185324	0.9999185976684842	1.0888750081142597E-4	41163
9	989168098	262144	0	0.9916273138845204	0.999799425723789	4.187537814854466E-4	62536
10	989168064	1048576	5	0.983829739776152	0.9995254239872715	0.0016541119914045867	11956
11	989168031	4194304	4960	0.9684790591458167	0.9989814991022692	0.008582437729916246	1084
12	989167998	16777216	426439	0.9386836380446671	0.9980695692458524	0.1181273118549536	498
13	989167965	67108864	9489190	0.8817244379724731	0.996697181705177	1.3226291369998282	312
14	989167932	268435456	99956562	0.7785257094242315	0.9947109218256579	7.361331644443221	224
15	989167899	1073741824	683982908	0.6122562535766236	0.9918955894391189	29.219606650628037	186

Table A.36: *Gallus gallus*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	12156679	4	0	0.998614835515522	0.9999087287670914	6.226194046674594E-6	4
2	12156662	16	0	0.9981638874223862	0.9999157103851467	1.3524086882388988E-5	16
3	12156645	64	0	0.9974472397606412	0.9999089673323568	2.5159941369584485E-5	64
4	12156628	256	0	0.9965661530483617	0.9998681099051334	5.2847233432004274E-5	254
5	12156611	1024	0	0.9943759819245677	0.9997428732031112	1.4937678370176966E-4	768
6	12156594	4096	0	0.9896817315771177	0.9993146519168525	5.39239802277219E-4	506
7	12156577	16384	0	0.9799161392224144	0.9980105638174219	0.0020660434953663994	30
8	12156560	65536	0	0.9611450936778168	0.9940598641246542	0.008021976391588038	4
9	12156543	262144	93	0.9239483626224989	0.9824229559909564	0.037110884700478415	2
10	12156526	1048576	31338	0.8516684783136235	0.9516606344845808	0.8753380537293544	2
11	12156509	4194304	1005544	0.7175438277551557	0.8843428951236821	11.995297869137618	0
12	12156492	16777216	10169697	0.510983843036297	0.7762544333239255	49.63218724404279	0
13	12156475	67108864	57719272	0.294272805233425	0.6560543083295398	95.4337428413997	0
14	12156458	268435456	257664213	0.1511102164791751	0.5564911079127316	123.14779116032122	0
15	12156440	1073741824	1062446343	0.08554083267798795	0.4860511002177838	134.53232106858349	0

Table A.37: *Saccharomyces cerevisiae*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	12490080	4	0	0.9993018459449419	0.999978818924562	1.8667000307340982E-6	4
2	12490077	16	0	0.998949005678668	0.9999746283929273	6.262413245586477E-6	16
3	12490074	64	0	0.9980115410044809	0.9999527520561629	1.819060462025271E-5	64
4	12490071	256	0	0.9970126671017322	0.9999137149635441	4.5537450376205474E-5	254
5	12490068	1024	0	0.9947005893002344	0.9997992946797678	1.411714062717796E-4	744
6	12490065	4096	0	0.9899840393144471	0.999455990569316	5.39756051650717E-4	548
7	12490062	16384	0	0.9804322828821826	0.9983941286049767	0.0021133193182477114	52
8	12490059	65536	0	0.9624729554920437	0.9951378406774212	0.008159856246518764	0
9	12490056	262144	547	0.9264839164852423	0.9852694197689916	0.04856892215175659	0
10	12490053	1048576	48496	0.8572665784524693	0.9561208118487056	0.99761262010086	0
11	12490050	4194304	1106811	0.7292910756962543	0.8801121851117107	11.1493143368679	0
12	12490047	16777216	10289340	0.5305365944579712	0.729988873397267	45.387463246036745	0
13	12490044	67108864	57601859	0.3133913699583444	0.5380776297508584	90.61505452786953	0
14	12490041	268435456	257271283	0.15769003480452948	0.3803339158805463	121.35366356982095	0
15	12490038	1073741824	1061902649	0.07999799520225637	0.2815302283902138	135.2822887694871	0

Table A.38: *Schizosaccharomyces pombe*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	1273397437	4	0	0.999783702250376	0.9999983903104772	1.4293694506832008E-7	4
2	1273397412	16	0	0.9997805469075353	0.9999980700914145	3.0263468152923645E-7	16
3	1273397387	64	0	0.9996361159487757	0.9999977469283051	6.199972396463768E-7	64
4	1273397362	256	0	0.9994268772483919	0.9999968209413015	1.5766540174357201E-6	256
5	1273397337	1024	0	0.9990357330235252	0.9999946051850798	4.553694749510483E-6	1024
6	1273397312	4096	0	0.998361013502752	0.9999892434153459	1.5087164037864443E-5	4094
7	1273397287	16384	0	0.9971086266339753	0.9999777422899198	5.1979653087018545E-5	15870
8	1273397262	65536	0	0.9948672804669498	0.999959366183067	1.7949865608850107E-4	45161
9	1273397237	262144	0	0.9907625690882507	0.9999402055431318	6.112082982750127E-4	60698
10	1273397212	1048576	0	0.983237445630594	0.999926898028833	0.0020253543117882484 7	28356
11	1273397187	4194304	97	0.9693274719005642	0.9999187469616712	0.007072274380812756	8616
12	1273397162	16777216	118061	0.9436641802410425	0.9999134248560033	0.056666968417778484	3580
13	1273397137	67108864	6606463	0.896914922151266	0.9999094832962127	0.9764112366767395	2282
14	1273397112	268435456	96686300	0.8168293497747465	0.9999062328417861	6.785374080696858	1765
15	1273397087	1073741824	719259928	0.6999256611303997	0.9999033148884986	23.186628423003306	1432

Table A.39: *D rerio*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	351195798	4	0	0.9998039896821317	0.999981400568275	2.0274214868227086E-7	4
2	351195797	16	0	0.9997584794558347	0.999991315960541	4.535914220084498E-7	16
3	351195796	64	0	0.9995254555951462	0.9999916017683124	1.0670599534705152E-6	64
4	351195795	256	0	0.999224429779975	0.9999888509054577	2.609121064220855E-6	256
5	351195794	1024	0	0.9987091758849481	0.9999788877443369	7.907813164983923E-6	1024
6	351195793	4096	0	0.9977497908125568	0.9999535599167729	2.6126552949055294E-5	4092
7	351195792	16384	0	0.9958067549966544	0.9998962400783803	9.012031301826948E-5	14530
8	351195791	65536	0	0.9920341300445711	0.9997767121263921	3.310785383420964E-4	21941
9	351195790	262144	0	0.984564945952228	0.9995968027845482	0.00126724351926724	4130
10	351195789	1048576	0	0.970077798968142	0.9994040766815275	0.004922561466583476	321
11	351195788	4194304	698	0.9420299539583317	0.9992349664658916	0.02043930934973679	128
12	351195787	16777216	246199	0.8880171731672852	0.9990838356682096	0.2935113764826776	80
13	351195786	67108864	10322639	0.786881628471476	0.9989378703139917	5.016088197704734	64
14	351195785	268435456	129811603	0.6185830390874423	0.9987839672916633	28.60404173064644	58
15	351195784	1073741824	845772629	0.40885130898951794	0.9986181842929254	71.70552774921185	46

Table A.40: *Fugu*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	118960067	4	0	0.999448125731133	0.9999911623814948	8.856814132391886E-7	4
2	118960062	16	0	0.9993725036895156	0.9999898911400662	2.291060317364609E-6	16
3	118960057	64	0	0.998862029798792	0.9999816639591258	7.104354870412616E-6	64
4	118960052	256	0	0.9984513540730463	0.9999658236553447	2.0463569426906185E-5	256
5	118960047	1024	0	0.9971760518890851	0.99992181802441	6.043087081783871E-5	1012
6	118960042	4096	0	0.9951630481098855	0.9998316516698256	1.6716847346917075E-4	3540
7	118960037	16384	0	0.9914747924969122	0.9995987856268155	4.968560272028137E-4	6204
8	118960032	65536	0	0.9848183884146904	0.9989885848468879	0.0015625814119903363	3162
9	118960027	262144	0	0.9721778728244572	0.9973971160900452	0.005163457798287983	422
10	118960022	1048576	757	0.9482676121226675	0.9934645140179809	0.02005625885594221	60
11	118960017	4194304	81726	0.9026679947431413	0.9844209576462167	0.23399621347382343	12
12	118960012	16777216	2524269	0.8177054235670387	0.9670047857038383	3.370989558232293	8
13	118960007	67108864	30540786	0.6737377545715847	0.9403816451645939	20.012987465819464	6
14	118960002	268435456	203137567	0.48121947745091664	0.9072193453216522	55.4972007003932	4
15	118959997	1073741824	986006688	0.2992811776886646	0.8708188059880304	92.64419451446601	4

Table A.41: *Arabidopsis thaliana*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	350945253	4	0	0.9998448675412059	0.9999970733493013	7.309407809977072E-8	4
2	350945241	16	0	0.9998264487079909	0.9999975319814426	1.7660520880101568E-7	16
3	350945229	64	0	0.9996674267368371	0.9999963107379024	4.804258264173293E-7	64
4	350945217	256	0	0.999310655942064	0.999989337465792	2.3255346293493246E-6	256
5	350945205	1024	0	0.9986183626586378	0.9999718794343799	9.74968994503448E-6	1024
6	350945193	4096	0	0.9974588653220277	0.9999203584808606	4.0536911191870837E-5	4070
7	350945181	16384	0	0.9952539624699961	0.9997881603620102	1.4976009536899208E-4	15166
8	350945169	65536	0	0.9912862171355321	0.9994305699508386	5.371691319903114E-4	21232
9	350945157	262144	0	0.9836697418793559	0.998556738890245	0.0017690079440110433	4042
10	350945145	1048576	0	0.9691250693894056	0.9967429858066764	0.0058303179776666445	667
11	350945133	4194304	40	0.9413727586870395	0.9938446328841647	0.020656968040061785	148
12	350945121	16777216	148755	0.8891377008201804	0.9905100982303576	0.24988823231813298	48
13	350945109	67108864	10813095	0.7941678195606368	0.9876158738212487	5.670251005859593	16
14	350945097	268435456	142499241	0.6461594959966059	0.9853386570205307	28.94450651099929	5
15	350945085	1073741824	884021306	0.48443597949234707	0.9835295872792151	61.20249528361909	4

Table A.42: *Oryza sativa*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	427311570	4	0	0.9992383730681572	0.9999877890976763	1.6741220983355467E-6	4
2	427311569	16	0	0.9991894509179554	0.9999873394830076	3.1376832358253446E-6	16
3	427311568	64	0	0.9988650435974156	0.9999866023495367	4.81210440616671E-6	64
4	427311567	256	0	0.998692179095634	0.9999864184159882	7.452185210389043E-6	256
5	427311566	1024	0	0.9982522682290327	0.9999840694974902	1.4533401265843069E-5	1024
6	427311565	4096	0	0.997470557577818	0.999977096243076	3.7547146525927234E-5	4018
7	427311564	16384	0	0.9958334195701757	0.9999596443819321	1.1335584371389709E-4	12796
8	427311563	65536	0	0.9926253224277949	0.9999161493436517	3.672442246990815E-4	21511
9	427311562	262144	0	0.9864166371421516	0.999810149499812	0.0012495675762495939	11882
10	427311561	1048576	0	0.9744270480900937	0.9995717983432942	0.004447524797871407	3688
11	427311560	4194304	4217	0.951285067036333	0.9991126636216213	0.021630160104631934	1472
12	427311559	16777216	706351	0.9074888259692502	0.9984181874017242	0.44551994683033763	783
13	427311558	67108864	16337462	0.8282776193945122	0.9975944612699182	4.495985883376127	374
14	427311557	268435456	151243248	0.7040732717650321	0.9967772046447246	19.397588894032364	162
15	427311556	1073741824	877897293	0.5527645313668981	0.9960163648554697	46.032315330428325	68

Table A.43: *Populus trichocarpa*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	327855916	4	0	0.9994710603300506	0.999993145633198	8.210719198878467E-7	4
2	327855896	16	0	0.9994299050214427	0.9999938503415968	1.56694291554258E-6	16
3	327855876	64	0	0.9991723009411612	0.9999930157447794	2.7331091587523005E-6	64
4	327855856	256	0	0.998925814520147	0.9999917721470439	5.035002474417155E-6	256
5	327855836	1024	0	0.9984744941371121	0.9999881209212635	1.1266517207755956E-5	1024
6	327855816	4096	0	0.9975528846497571	0.9999783504949058	3.254448319386536E-5	3987
7	327855796	16384	0	0.995675043670724	0.9999525174382329	1.0969450938932744E-4	11364
8	327855776	65536	0	0.9920437546294747	0.9998841426825703	3.940726164068135E-4	17434
9	327855756	262144	0	0.9848496910330286	0.9997146294495589	0.0014638622341576802	7294
10	327855736	1048576	28	0.9711196878373358	0.9993444241878062	0.005617559885397787	1917
11	327855716	4194304	29113	0.9447614572014965	0.9986290523907491	0.05089758747402787	674
12	327855696	16777216	1505840	0.8954349049955197	0.9975253693421472	0.8656568723344017	239
13	327855676	67108864	22026986	0.8085567809416239	0.9960963303318023	6.124280075598392	90
14	327855656	268435456	169196096	0.6761086500822788	0.9944885988660987	23.299400159547012	39
15	327855636	1073741824	913774764	0.5181619632123695	0.9927754273731368	52.346780855605665	24

Table A.44: *Vitis vinifera*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	33976983	4	0	0.9991711742034306	0.9999954802898957	2.9750360718926753E-6	4
2	33976975	16	0	0.9989016091044008	0.9999941928582962	7.252405744633167E-6	16
3	33976967	64	0	0.9984061849899669	0.9999894655307074	1.7808279202033126E-5	64
4	33976959	256	0	0.9976602379277086	0.9999848630318594	3.9006495510187825E-5	249
5	33976951	1024	0	0.9962030436456761	0.99997682950472	9.644041618827295E-5	670
6	33976943	4096	0	0.9938872664324157	0.9999646676520253	2.630135459882829E-4	1182
7	33976935	16384	1	0.9895001417873626	0.9999470471512443	8.912180936667874E-4	1074
8	33976927	65536	375	0.9814374913893773	0.9999251274205605	0.005675937627359039	609
9	33976919	262144	14480	0.966348832276405	0.9999009314278028	0.06957692302175984	382
10	33976911	1048576	214942	0.938972939594185	0.9998738896702273	0.5815321554331817	229
11	33976903	4194304	1869928	0.8910444250907742	0.999841352027168	2.925376062607574	164
12	33976895	16777216	11557283	0.8156136103666918	0.9997974329959213	9.844713423709308	123
13	33976887	67108864	57607897	0.7121946162990153	0.9997404842714959	23.779167836949036	106
14	33976879	268435456	254110448	0.5936852822768095	0.9996760381616496	43.48529533961111	87
15	33976871	1073741824	1055071824	0.4795927205892503	0.9995949872738686	64.28908473791498	72

Table A.45: *Dictyostelium discoideum*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	32110796	4	0	0.9983690220572545	0.9998126792770597	8.588131538643408E-6	4
2	32110760	16	0	0.9978215713362125	0.9997802626442887	2.7709273025797254E-5	16
3	32110724	64	0	0.9958596386677547	0.9997196253848853	6.798407618759075E-5	64
4	32110688	256	0	0.9946523413014383	0.9995993281461522	1.4815034104203142E-4	250
5	32110652	1024	0	0.992364714363321	0.9992652361172537	3.7218063869694636E-4	868
6	32110616	4096	0	0.988987878650475	0.9984228820179669	9.342363203882666E-4	1694
7	32110580	16384	0	0.982622487666059	0.9964391658312742	0.002334732152245713	730
8	32110544	65536	0	0.9704004080404244	0.9917409261881082	0.006469896959715569	52
9	32110508	262144	38	0.9468395828555562	0.9815777049573077	0.02002391863860716	8
10	32110472	1048576	12298	0.9019812913369819	0.9637636172208268	0.18372865663864393	6
11	32110436	4194304	514871	0.8172990861911685	0.9406587526618867	2.8929685349820167	4
12	32110400	16777216	6971636	0.6705159387612736	0.9162124814389421	19.014904586469797	3
13	32110364	67108864	49106129	0.4677637413266321	0.8936147522418688	56.738916216842284	0
14	32110328	268435456	244042648	0.2719609092750469	0.8731318984277123	97.44024570335904	0
15	32110291	1073741824	1045987466	0.14258419520396126	0.8546505054729854	122.9101094000992	0

Table A.46: *Leishmania infantum*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	22859235	4	0	0.9997919002976259	0.9999996889676223	4.1278385917168536E-7	4
2	22859220	16	0	0.9992106029864536	0.9999976628121167	6.88974960170454E-6	16
3	22859205	64	0	0.9974229637469895	0.9999868855448468	3.858003468246238E-5	64
4	22859190	256	0	0.9962577414160344	0.9999800046893043	1.1941297195870779E-4	195
5	22859174	1024	0	0.9941767799658903	0.999969228216889	3.245653988197988E-4	450
6	22859159	4096	0	0.9913680551414862	0.999956711928505	7.925512943417197E-4	658
7	22859144	16384	0	0.9861145281730584	0.9999394969410175	0.002135371292923203	662
8	22859129	65536	498	0.9763931075414116	0.9999133504624335	0.012127906799793706	468
9	22859114	262144	19742	0.9585300637636262	0.9998821412771095	0.14713466036893136	328
10	22859099	1048576	276381	0.9265188448591084	0.9998455763282068	1.0842464186166012	193
11	22859084	4194304	2227526	0.873049768748389	0.9998101791429982	4.719302608009409	102
12	22859069	16777216	12789982	0.793680398794894	0.9997743206797464	13.372857273425248	66
13	22859054	67108864	60436828	0.6911092646266115	0.9997352746684373	27.82851762726649	56
14	22859039	268435456	258817417	0.5766194720609208	0.9996900153664835	46.31413849142063	44
15	22859024	1073741824	1061349856	0.46427529014362123	0.9996376331065303	65.87260383982937	28

Table A.47: *Plasmodium falciparum*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	27127879	4	0	0.9968198398407778	0.9942312541013579	4.079487264503091E-5	4
2	27127866	16	0	0.9959112154269709	0.9978051929573346	7.551560927696623E-5	16
3	27127853	64	0	0.9945934165892155	0.998446448101724	1.221343080475803E-4	64
4	27127840	256	0	0.9933752926882494	0.9986995594635447	1.9118172145406133E-4	256
5	27127827	1024	0	0.9906851735673484	0.9984639896632745	3.9213885100670065E-4	858
6	27127814	4096	0	0.9862760412615628	0.9976004875650063	0.0010055535653043463	1317
7	27127801	16384	0	0.9776914833605569	0.9952491874108912	0.003108018974619332	132
8	27127788	65536	0	0.9614620624431303	0.989722519320648	0.009869122368861227	21
9	27127775	262144	0	0.9300615328754386	0.9780735716110013	0.031034483398859863	10
10	27127762	1048576	970	0.8709294559573326	0.9576295289145722	0.11911367055672793	6
11	27127749	4194304	211579	0.7609816059563217	0.9286229704460202	2.810996961381792	4
12	27127736	16777216	6040788	0.5755922646843805	0.8946508242492681	30.21665189886568	2
13	27127723	67108864	49676990	0.3496336201899437	0.8584921404476974	80.4284564194154	2
14	27127710	268435456	247486401	0.18925980114060492	0.8210122260712271	113.6976952776606	2
15	27127697	1073741824	1051371384	0.11392157616623333	0.7797335500195864	127.93931084485777	2

Table A.48: *Trypanosoma brucei*, complete genome results for N oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	226204033	4	0	0.999509526870372	0.9999716382921379	9.607006394345006E-7	4
2	226204032	16	0	0.9993153747144525	0.9999838564616232	1.8185857756619128E-6	16
3	226204031	64	0	0.9991784098666217	0.9999866013066289	2.8011671076356302E-6	64
4	226204030	256	0	0.9989345813158148	0.9999847964177657	5.1887004363544E-6	256
5	226204029	1024	0	0.998398503326393	0.9999772584875114	1.2498772309070501E-5	1024
6	226204028	4096	0	0.9973582875367719	0.999956795601877	3.769913532803876E-5	3655
7	226204027	16384	0	0.9951526901861919	0.9999019081328303	1.386823021680846E-4	10510
8	226204026	65536	0	0.9908999320816686	0.9997799863311405	5.296400083560931E-4	12967
9	226204025	262144	1	0.9827676143251651	0.999571197089791	0.0020068170172642423	2416
10	226204024	1048576	2685	0.967239336113667	0.9992942889980457	0.011832073692897406	988
11	226204023	4194304	148896	0.9375033617328724	0.9989694982771051	0.1432737863941503	746
12	226204022	16777216	2826013	0.8816549601403639	0.9985967206073304	1.5326555913307238	610
13	226204021	67108864	27115685	0.7828428478731596	0.9981548881993155	7.561390369218011	542
14	226204020	268435456	179693850	0.6289620007637353	0.9976139775875927	30.833165983239873	480
15	226204019	1073741824	936677941	0.4456636731993696	0.9969436020826898	68.84794014307099	422
16	226204018	4294967296	4129284501	0.30377320707008837	0.9961190970674301	96.73609582499617	356
17	226204017	17179869184	17000687316	0.22588551113130761	0.9951256186083093	110.69768186932484	318
18	226204016	68719476736	68533860144	0.188118888216379	0.9939213327986125	117.1002558760231	266
19	226204015	274877906944	274688656170	0.16834072551718415	0.9924363802913533	120.36734214842866	202
20	226204014	1099511627776	1099319844281	0.1558301171437214	0.9906142161584739	122.4205699277329	160
21	226204013	4398046511104	4397852673504	0.14633773981719767	0.988358708103284	123.97661385967967	118
22	226204012	17592186044416	17591990406073	0.13829486366492916	0.9856469658859505	125.296405442341	86
23	226204011	70368744177664	70368546901651	0.1310944039803078	0.9825012939688139	126.47883611834715	64
24	226204010	281474976710656	281474777919447	0.12447973844495508	0.9788859484135433	127.56541212628719	46
25	226204009	1,13E+015	1,13E+015	0.11833911396327201	0.9746452826232856	128.57672238309553	36
26	226204008	4,50E+015	4,50E+015	0.1126115280857446	0.9695975701236627	129.52180054910835	30
27	226204007	1,80E+016	1,80E+016	0.10724691539173314	0.963554168866376	130.40826915326733	20
28	226204006	7,21E+016	7,21E+016	0.10220351269994754	0.9565179325553471	131.24437409576703	16
29	226204005	2,88E+017	2,88E+017	0.09745057343259678	0.9485375616921597	132.0341378604803	8
30	226204004	1,15E+018	1,15E+018	0.09296944186717404	0.9391975156469271	132.78142978745007	6

Table A.49: *Homo sapiens*, results for chromosome 1, oligonucleotide order 1 to 30.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	237895365	4	0	0.9992742397482187	0.9999649060567204	1.6843199789095115E-6	4
2	237895364	16	0	0.9992251719541706	0.9999742118632954	3.241990470003755E-6	16
3	237895363	64	0	0.9989020845269692	0.9999768895783635	5.1174977173980324E-6	64
4	237895362	256	0	0.9986430546720789	0.9999752568428498	8.386997436250443E-6	256
5	237895361	1024	0	0.9982012385689185	0.9999689214778062	1.59623566389757E-5	1020
6	237895360	4096	0	0.9971634755717808	0.9999503496023244	4.232295455583977E-5	3673
7	237895359	16384	0	0.9950771843346469	0.9998969316511188	1.4129300924666175E-4	10658
8	237895358	65536	0	0.9910149528852934	0.9997684637801943	5.08320049496274E-4	13967
9	237895357	262144	5	0.9831655941061515	0.9995202389222985	0.0018912015212594992	2862
10	237895356	1048576	3250	0.9680870441203568	0.9991387831707149	0.011788953555222347	1022
11	237895355	4194304	157690	0.939299289807487	0.9986591751855417	0.1351457473057481	722
12	237895354	16777216	2816890	0.8853176090189638	0.9981103709856968	1.420159385804014	589
13	237895353	67108864	26827328	0.7891030137104023	0.9974768013048787	6.988005691662479	494
14	237895352	268435456	177113019	0.6371561895837292	0.9967176980865999	28.916028122375124	410
15	237895351	1073741824	929073405	0.4501524201706657	0.9957854950228469	67.40346934760026	336

Table A.50: *Homo sapiens*, results for chromosome 2, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	195304078	4	0	0.9996676669495862	0.9999934795370717	3.484013263095536E-7	4
2	195304077	16	0	0.9996125477708282	0.9999945700975681	7.664783030694887E-7	16
3	195304076	64	0	0.9993764902274748	0.9999918775514739	1.7577555305960344E-6	64
4	195304075	256	0	0.9989748652197861	0.9999850225411078	5.152355548668387E-6	2569
5	195304074	1024	0	0.998260466394572	0.999972775819907	1.4194870396512121E-5	1022
6	195304073	4096	0	0.997060486291036	0.9999407581968358	4.626266020551152E-5	3483
7	195304072	16384	0	0.9946251095061653	0.9998639997121029	1.6208207342701045E-4	10016
8	195304071	65536	0	0.9901389203505133	0.9996812603594789	5.980936992627774E-4	9964
9	195304070	262144	9	0.9815595343199965	0.999331913473119	0.0022935514940381364	1692
10	195304069	1048576	6484	0.9652063931141137	0.9988051918611306	0.0175985594010713	505
11	195304068	4194304	224341	0.9338082911821376	0.9981505633946866	0.20391805577671618	278
12	195304067	16777216	3486798	0.8750996209413294	0.9974202656296255	1.8765789414750673	214
13	195304066	67108864	29711400	0.7718851076044674	0.9966067428925117	8.444830017804158	178
14	195304065	268435456	187160239	0.611849543428602	0.9956508692278538	33.785884751710554	136
15	195304064	1073741824	950361924	0.4237791692854891	0.994505569492175	72.80574497232547	102

Table A.51: *Homo sapiens*, results for chromosome 3, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	187939711	4	0	0.999756789026881	0.9999971043699054	1.9406440932657306E-7	4
2	187939710	16	0	0.9996750447257794	0.9999974700132749	5.057284226969091E-7	16
3	187939709	64	0	0.9994601300569216	0.9999958114750329	1.6240565773862229E-6	64
4	187939708	256	0	0.9990081127507126	0.99999099411105	5.020138218844636E-6	256
5	187939707	1024	0	0.998325968444763	0.9999800506181059	1.517105874817326E-5	1020
6	187939706	4096	0	0.9971798348987521	0.9999519737331773	4.956806261571905E-5	3447
7	187939705	16384	0	0.994862080899829	0.9998777071315369	1.7423770912899749E-4	9878
8	187939704	65536	0	0.9903897954420531	0.9996934874080883	6.396210146477338E-4	9527
9	187939703	262144	25	0.9815931229815767	0.9992974323688417	0.0024812750865450916	1918
10	187939702	1048576	7796	0.9649120173660806	0.998621873187339	0.020067796240297035	535
11	187939701	4194304	248132	0.9331142545555077	0.9977174551065642	0.23139772677013482	312
12	187939700	16777216	3734006	0.8738556675359171	0.996664991303901	2.003929078573635	234
13	187939699	67108864	30832372	0.7703934654061567	0.9954867835876487	8.872944715100386	180
14	187939698	268435456	190977349	0.6116219469502393	0.9941293763907589	34.090943899429206	162
15	187939697	1073741824	956652787	0.42679783611654964	0.992510203644173	71.83023871192164	144

Table A.52: *Homo sapiens*, results for chromosome 4, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	177846453	4	0	0.9990530314371803	0.9999476762708609	2.8664396562469263E-6	4
2	177846452	16	0	0.9989625826215527	0.9999601523442087	5.830397421012955E-6	16
3	177846451	64	0	0.9985590547432402	0.9999628074444289	9.20034594687525E-6	64
4	177846450	256	0	0.9982787117763666	0.9999596464515558	1.4201516631745866E-5	256
5	177846449	1024	0	0.9976592223103651	0.9999503828559617	2.4843226330418866E-5	1018
6	177846448	4096	0	0.9966687217728408	0.9999260404513721	5.680091358312161E-5	3454
7	177846447	16384	0	0.9943340504294697	0.9998600407221218	1.7710375089927723E-4	9690
8	177846446	65536	0	0.9896954758376223	0.9996926903965481	6.443478613502773E-4	8112
9	177846445	262144	20	0.9807160666045363	0.9993487760498134	0.002526381423291092	1462
10	177846444	1048576	7015	0.9634648247451043	0.998790206766109	0.02038724525523418	556
11	177846443	4194304	237196	0.9305083487106909	0.9980715771709048	0.24040551535382787	374
12	177846442	16777216	3669825	0.869151292888952	0.9972502907014599	2.1135734896821603	280
13	177846441	67108864	30720580	0.7621720133269352	0.996311967054924	9.513273839185636	232
14	177846440	268435456	191691461	0.5992311513235801	0.9952014951272491	36.38706403039634	188
15	177846439	1073741824	959986529	0.41397005424438105	0.9938466939893817	74.70563147545458	158

Table A.53: *Homo sapiens*, results for chromosome 5, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	169096731	4	0	0.9998125510776432	0.9999973948726539	2.0327028895072564E-7	4
2	169096730	16	0	0.999655090905661	0.999996779411082	5.662299134607765E-7	16
3	169096729	64	0	0.9994333125154656	0.9999956059393699	1.3803018454929545E-6	64
4	169096728	256	0	0.9991878080574096	0.9999924861031002	3.5123223163159122E-6	256
5	169096727	1024	0	0.9984920287664705	0.9999836706880455	1.1665255641993168E-5	1018
6	169096726	4096	0	0.9971588095679629	0.9999571473268452	4.400150889926857E-5	3426
7	169096725	16384	0	0.9945668788085635	0.9998867926730636	1.6986720580248506E-4	9550
8	169096724	65536	0	0.9896504559130311	0.9997156924027479	6.573679513001367E-4	7430
9	169096723	262144	18	0.9801928450145069	0.999368576523686	0.002603020085067834	1392
10	169096722	1048576	6375	0.9621166458803383	0.9988095370812108	0.02066554801731579	598
11	169096721	4194304	231935	0.927962500230859	0.9981044281021358	0.25546546727101616	418
12	169096720	16777216	3686690	0.8643887415438928	0.99729097965294	2.270481848233169	314
13	169096719	67108864	31052286	0.7540525372346225	0.9963582028919785	10.33035587639768	256
14	169096718	268435456	193675263	0.5875545970087959	0.9952423824472815	38.448480674592375	208
15	169096717	1073741824	964347244	0.40188563802808774	0.993888239257891	76.908640533985	162

Table A.54: *Homo sapiens*, results for chromosome 6, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	155401272	4	0	0.999478884574381	0.9999826389766462	7.950648636854843E-7	4
2	155401271	16	0	0.9993670064641877	0.9999781455160364	2.3793586292275035E-6	16
3	155401270	64	0	0.9989873699230386	0.9999713557680774	4.998993667891159E-6	64
4	155401269	256	0	0.9985491752966316	0.9999587006171047	1.1197865309942118E-5	256
5	155401268	1024	0	0.9976999930270839	0.9999371552135822	2.8844702197556138E-5	1008
6	155401267	4096	0	0.9962525466410772	0.9998837574088413	8.551522253569526E-5	3395
7	155401266	16384	0	0.9935657538336914	0.9997665691891957	2.700497862720481E-4	9096
8	155401265	65536	0	0.9884107635803351	0.9995297306324703	9.119156428381595E-4	5720
9	155401264	262144	16	0.9786485649177217	0.9991528818466753	0.003264564499055661	1140
10	155401263	1048576	6058	0.9600988378067429	0.9986639262832984	0.022288587332406663	455
11	155401262	4194304	228523	0.9247299291559163	0.9981171490506007	0.2784408589499148	292
12	155401261	16777216	3707829	0.85916006177067	0.9975090611395593	2.4386186388016684	242
13	155401260	67108864	31588857	0.7464471394890878	0.9968076399555671	11.477164700354443	214
14	155401259	268435456	197330628	0.5805267703783532	0.9959629336826754	40.63987099932105	176
15	155401258	1073741824	972669890	0.40425518305649755	0.9949261793985229	77.15111171401742	136

Table A.55: *Homo sapiens*, results for chromosome 7, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	143199949	4	0	0.999611599023684	0.9999855052342844	6.758412674942734E-7	4
2	143199948	16	0	0.9993962846969748	0.9999901312804782	1.396569465585963E-6	16
3	143199947	64	0	0.9991617664495365	0.9999892635044698	2.966255286310379E-6	64
4	143199946	256	0	0.9987824576414296	0.9999827441158509	7.053823534279326E-6	256
5	143199945	1024	0	0.9979787631901674	0.9999666519339231	2.0283146126323077E-5	1018
6	143199944	4096	0	0.9964957947190258	0.999922715705542	6.738866955928523E-5	3324
7	143199943	16384	0	0.9935468060905583	0.9998090939323014	2.502893245406882E-4	8744
8	143199942	65536	0	0.9882449533394364	0.9995485105599582	9.138148037444563E-4	5142
9	143199941	262144	21	0.9780454727980649	0.9990525643412209	0.003435039954207958	946
10	143199940	1048576	8202	0.9585179225633754	0.9983051167285111	0.027071598728229136	403
11	143199939	4194304	265629	0.9213509651006206	0.9974006904677573	0.33111301693346495	278
12	143199938	16777216	3978973	0.8525478342036712	0.9964151987482253	2.7105326980698528	212
13	143199937	67108864	32544544	0.7343143454036576	0.995298504734097	12.480680116904344	166
14	143199936	268435456	199832543	0.5603425968011606	0.993956651666725	43.76059349130214	144
15	143199935	1073741824	977014014	0.37612000312709637	0.9923105992873309	82.01279127143792	106

Table A.56: *Homo sapiens*, results for chromosome 8, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	120983611	4	0	0.9997203505522744	0.9999947415245565	2.261280634578355E-7	4
2	120983610	16	0	0.9995835551609016	0.9999921249386614	8.977235844012584E-7	16
3	120983609	64	0	0.9992766540796448	0.9999895286588526	2.4058326989433615E-6	64
4	120983608	256	0	0.9988264443229367	0.9999827233846352	7.012927321026454E-6	256
5	120983607	1024	0	0.9979221234493364	0.9999660608224521	2.164475300348873E-5	1008
6	120983606	4096	0	0.9964398730188286	0.9999238897011195	7.14415294018184E-5	3258
7	120983605	16384	0	0.9932845033010878	0.9998087159435131	2.5974775943196036E-4	8058
8	120983604	65536	0	0.987574051769858	0.9995565511697071	9.598942632039954E-4	3361
9	120983603	262144	48	0.9764986251897292	0.9990920173684489	0.0038343820275959766	634
10	120983602	1048576	9723	0.9554999197329238	0.9984469599467118	0.03253525417525305	333
11	120983601	4194304	303205	0.9157508049376047	0.9976952336158558	0.4321269674820538	226
12	120983600	16777216	4361602	0.8426382170806621	0.9968466090141763	3.19959416202838	154
13	120983599	67108864	34733828	0.7197858446912295	0.9958616004500619	14.961270460482941	120
14	120983598	268435456	208344741	0.5480046725011435	0.9946675148237467	47.21553501048319	90
15	120983597	1073741824	993179051	0.3829837196855703	0.9932166873615403	80.70094446810404	62

Table A.57: *Homo sapiens*, results for chromosome 9, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	131735771	4	0	0.9996301308321185	0.9999866907552822	4.755836261715953E-7	4
2	131735770	16	0	0.9994770213131938	0.9999918874865495	1.0053332104821508E-6	16
3	131735769	64	0	0.9992765897924049	0.999990840028135	2.1939136365382816E-6	64
4	131735768	256	0	0.9988524452979239	0.9999836383436556	6.329573319075902E-6	256
5	131735767	1024	0	0.998055858284865	0.9999679421240095	2.095775521440313E-5	1002
6	131735766	4096	0	0.996631127495019	0.9999279230370902	7.226589594186519E-5	3331
7	131735765	16384	0	0.9936801596741781	0.9998299645941304	2.5552847480909925E-4	8568
8	131735764	65536	0	0.9880913887590921	0.9996088348625369	9.108921985871635E-4	4049
9	131735763	262144	26	0.977263645559938	0.9992132504416978	0.0035045120715240237	814
10	131735762	1048576	7584	0.9565258900616523	0.9986504140224663	0.027966339987320517	425
11	131735761	4194304	259586	0.91711779005854	0.9979906415188445	0.3647101174474956	320
12	131735760	16777216	3980667	0.8441221199164145	0.9972370196746682	3.0016581234693054	242
13	131735759	67108864	33055014	0.7197440142277542	0.9963519704210473	14.24366591281296	198
14	131735758	268435456	203085107	0.5414573315773534	0.9952735093689136	47.4025192357612	156
15	131735757	1073741824	984130334	0.36189298248007185	0.993926063235738	84.61819220026354	136

Table A.58: *Homo sapiens*, results for chromosome 10, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	131245847	4	0	0.999672926793638	0.9999908650581305	4.3428717691892575E-7	4
2	131245846	16	0	0.9995507972115171	0.999991662180991	1.27508581996755E-6	16
3	131245845	64	0	0.9993084657270483	0.9999892580209742	2.9066460775226566E-6	64
4	131245844	256	0	0.9988207702790193	0.9999822971711267	6.979264070013963E-6	256
5	131245843	1024	0	0.9979975213386377	0.9999643931269576	1.9718806562858098E-5	1014
6	131245842	4096	0	0.9965425190384317	0.9999206587290441	6.389151891551622E-5	3323
7	131245841	16384	0	0.9936571780586937	0.9998180470688447	2.2413373287472136E-4	8618
8	131245840	65536	0	0.9883257937927785	0.9995847644249087	8.158759586955373E-4	3990
9	131245839	262144	21	0.9778334839247742	0.9991440042354691	0.003236507340657776	766
10	131245838	1048576	8134	0.9576717091783131	0.9984887277131372	0.027880547693125837	391
11	131245837	4194304	269668	0.9190804581481697	0.997691831403153	0.37077345887829855	262
12	131245836	16777216	4068168	0.8474077760455577	0.9967884018096977	2.958865643551788	188
13	131245835	67108864	33147492	0.7250488977421645	0.9957475320072915	13.805736118509149	152
14	131245834	268435456	202817678	0.547953552567619	0.9944937515312556	46.782557824150786	130
15	131245833	1073741824	983446616	0.3678850360148196	0.9929668980504255	84.15416924811669	96

Table A.59: *Homo sapiens*, results for chromosome 11, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	130303032	4	0	0.9997143427944178	0.9999951628512523	2.397401963513336E-7	4
2	130303031	16	0	0.9996620032576219	0.9999948087994156	6.538482956212259E-7	16
3	130303030	64	0	0.9993798455799531	0.9999924794196559	1.7613200755564048E-6	64
4	130303029	256	0	0.9989871225480108	0.999985445951869	5.217604888676083E-6	256
5	130303028	1024	0	0.9981894357819528	0.9999718901881606	1.619483264116959E-5	1010
6	130303027	4096	0	0.9967136603818114	0.9999370743549314	5.6548334025498255E-5	3311
7	130303026	16384	0	0.9937779956084827	0.9998435608094602	2.1929710834297614E-4	8484
8	130303025	65536	0	0.9882458906844258	0.9996368476415542	8.361170859345895E-4	3974
9	130303024	262144	31	0.9774746286778425	0.9992727184881346	0.0034088957435275623	806
10	130303023	1048576	8784	0.9570641810819692	0.9987722767445708	0.029801518434314774	467
11	130303022	4194304	282580	0.9183009124684768	0.9981798916979996	0.3881122989654101	332
12	130303021	16777216	4184905	0.8468532283683584	0.9975063084681809	3.0286779495194813	248
13	130303020	67108864	33693867	0.7253119536293173	0.9967197184964763	14.003213802590816	194
14	130303019	268435456	204300520	0.5499565746822797	0.9957692637546383	46.73921183726261	168
15	130303018	1073741824	985578118	0.37151871647362766	0.9945998034222634	83.59881181439707	136

Table A.60: *Homo sapiens*, results for chromosome 12, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	95746838	4	0	0.9991297884949475	0.999950191393849	3.0864178357579317E-6	4
2	95746837	16	0	0.9988482021604537	0.9999575544730948	5.954540877635833E-6	16
3	95746836	64	0	0.9985571742548234	0.9999582212973994	1.0395677018454501E-5	64
4	95746835	256	0	0.9980671737086662	0.9999535338124586	1.7462213152635228E-5	256
5	95746834	1024	0	0.9972600242844584	0.9999383616734042	3.513958011446611E-5	992
6	95746833	4096	0	0.9958252091742815	0.9998963828059065	9.358511296835027E-5	3040
7	95746832	16384	0	0.992651579323272	0.999774304057346	3.273652143670194E-4	6542
8	95746831	65536	0	0.9864876676701707	0.9994735974084001	0.001206160567264153	2076
9	95746830	262144	161	0.9743842172111599	0.9988195197497766	0.0053574343280710845	428
10	95746829	1048576	18533	0.9509624804389083	0.9976792856691545	0.05662809751525289	211
11	95746828	4194304	430320	0.9063687415315732	0.9961349067743959	0.6861604483145851	148
12	95746827	16777216	5178960	0.8248972260981557	0.9943199615917262	4.213792613314185	124
13	95746826	67108864	37841328	0.6894130359997521	0.9922441339913709	18.94286791019905	98
14	95746825	268435456	215986304	0.5026144104517304	0.9898050642510368	55.24013579185492	78
15	95746824	1073741824	1004284452	0.32513792833483435	0.9868666837003696	91.3849007535818	68

Table A.61: *Homo sapiens*, results for chromosome 13, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	88290585	4	0	0.9971756105138504	0.9993229843292861	2.8013423745099378E-5	4
2	88290584	16	0	0.9968199893207185	0.9995330730769051	5.401774857395432E-5	16
3	88290583	64	0	0.9956503288691615	0.9995951799879723	8.068326718250527E-5	64
4	88290582	256	0	0.9951206347240977	0.999605844771679	1.1168497980921203E-4	256
5	88290581	1024	0	0.9941153972018827	0.9995964763119046	1.5377566343094254E-4	974
6	88290580	4096	0	0.9928011346170792	0.9995500809335517	2.4099683054439747E-4	2912
7	88290579	16384	0	0.9898943351589075	0.9994389293433618	4.992910507264986E-4	5058
8	88290578	65536	0	0.9838387398483222	0.9991662151661349	0.0014417319930556214	1048
9	88290577	262144	107	0.9715731951780087	0.9986702784988785	0.0055663328671489995	294
10	88290576	1048576	16026	0.9473706344378137	0.997937222256293	0.05769984656793952	173
11	88290575	4194304	415618	0.9007930234908992	0.9970268900548074	0.7468622151761499	126
12	88290574	16777216	5158040	0.8158850569937398	0.9959743109299969	4.618223245793526	92
13	88290573	67108864	38272279	0.6757063406984571	0.9947181527745321	21.19971992418424	80
14	88290572	268435456	218529225	0.48810054147117765	0.9931731893072594	59.13614154837592	64
15	88290571	1073741824	1009642107	0.32035130908826037	0.9912497388812749	93.22517400629309	48

Table A.62: *Homo sapiens*, results for chromosome 14, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	81920097	4	0	0.9994611700716126	0.9999758128384778	8.45792867054663E-7	4
2	81920096	16	0	0.9993594001647654	0.9999816792713327	2.2854224520610498E-6	16
3	81920095	64	0	0.9990237682219485	0.9999810131633157	4.4421572940627235E-6	64
4	81920094	256	0	0.9986428726510006	0.99997392520276	9.9756725114489E-6	256
5	81920093	1024	0	0.9976693507904099	0.999953991537517	2.794277606725228E-5	990
6	81920092	4096	0	0.995840226839589	0.9998931302490416	9.341005847933803E-5	3083
7	81920091	16384	0	0.9922041468435381	0.9997510247774092	3.4515305750631374E-4	5672
8	81920090	65536	0	0.9853610024110081	0.9994588497080171	0.0012996037728887101	1217
9	81920089	262144	112	0.9720071715254118	0.9989555143422316	0.0056899106640279	322
10	81920088	1048576	15972	0.9462763150352085	0.9982757098707965	0.059930498726988864	185
11	81920087	4194304	424448	0.8976384021662477	0.9974846531425727	0.8064860540134716	132
12	81920086	16777216	5241566	0.809341289021596	0.9965776900351533	4.993664142325823	108
13	81920085	67108864	39095767	0.6659730882847594	0.995510989635719	22.990800732600178	92
14	81920084	268435456	221481946	0.480882136302497	0.9942075952366198	60.84048275859557	72
15	81920083	1073741824	1014781035	0.3245452278167248	0.9925846749076006	92.31561128805143	56

Table A.63: *Homo sapiens*, results for chromosome 15, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	78990186	4	0	0.9977595950970416	0.9990429941830196	1.4576581663070214E-5	4
2	78990185	16	0	0.9976668367088899	0.9997089672629355	2.7494515718832244E-5	16
3	78990184	64	0	0.9967655221565251	0.9997801754230328	4.294306017905021E-5	64
4	78990183	256	0	0.9961824623193999	0.999800515034237	6.221127006754769E-5	256
5	78990182	1024	0	0.9954252542423564	0.999790403722004	9.733825524691041E-5	982
6	78990181	4096	0	0.9937491724446106	0.9997337876355227	1.9361203494411633E-4	2992
7	78990180	16384	0	0.990245445699706	0.9996025925329308	5.129400402077363E-4	4512
8	78990179	65536	0	0.9833365866913658	0.9993581694756832	0.0016322468146244267	988
9	78990178	262144	37	0.9696509609080765	0.9990017508515513	0.006057640163013232	430
10	78990177	1048576	10152	0.9433507662604681	0.9985787315158432	0.05344194421760484	293
11	78990176	4194304	353615	0.8930730069521556	0.9980751080176384	0.7829590383973588	228
12	78990175	16777216	4906551	0.8020531920583794	0.9974905132547598	5.497378089800741	188
13	78990174	67108864	39046778	0.6557703493601622	0.996773714410655	25.44285250994583	166
14	78990173	268435456	222939711	0.47407807297750826	0.9958731018786308	62.88246625466654	130
15	78990172	1073741824	1017661241	0.3275540152007771	0.9947341235478853	92.06551439454253	86

Table A.64: *Homo sapiens*, results for chromosome 16, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	79601053	4	0	0.9989675890342807	0.9997251165252407	3.0760039834758356E-6	4
2	79601052	16	0	0.9988715224517385	0.999936415768072	5.62990199492986E-6	16
3	79601051	64	0	0.9985141276589425	0.9999519777448574	9.308325322782464E-6	64
4	79601050	256	0	0.9980703143991191	0.999947351821217	1.686346637951109E-5	256
5	79601049	1024	0	0.9970195493277985	0.9999165185987579	4.090382127928819E-5	1006
6	79601048	4096	0	0.9952071234037019	0.9998426541467255	1.1846967426323037E-4	3146
7	79601047	16384	0	0.9914423864299172	0.9996962465231509	4.0346073466045786E-4	4926
8	79601046	65536	0	0.984299377171501	0.9994463704916443	0.0014835426165348458	1210
9	79601045	262144	40	0.9704448226778932	0.999129182900119	0.005978968791780862	538
10	79601044	1048576	9450	0.9441292378024589	0.9987897114752675	0.05216620974548507	353
11	79601043	4194304	350956	0.89441496388433	0.9984344541162241	0.7875674886331334	256
12	79601042	16777216	4916340	0.8044347962178686	0.9980273041857255	5.48569136119878	202
13	79601041	67108864	39215901	0.6606404808198425	0.9975385656978868	25.30007625385808	158
14	79601040	268435456	223430741	0.4834609196060755	0.9969365464549882	61.75940221114302	122
15	79601039	1073741824	1018300519	0.3414758694292922	0.99618690033315	90.19813243015349	92

Table A.65: *Homo sapiens*, results for chromosome 17, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	74660417	4	0	0.9993354845580357	0.9999788340838843	1.3396723223357027E-6	4
2	74660416	16	0	0.999269947812774	0.9999825103806027	3.077417458725957E-6	16
3	74660415	64	0	0.9988437647982535	0.9999770613750728	7.680200418743306E-6	64
4	74660414	256	0	0.9983269982939018	0.9999691900606551	1.501334995652779E-5	256
5	74660413	1024	0	0.9972459970185271	0.9999452570242585	3.727454358440506E-5	970
6	74660412	4096	0	0.9954664327327848	0.9998872822258381	1.1303163394150053E-4	2926
7	74660411	16384	0	0.9918117916602415	0.9997342182297372	3.903686797706442E-4	4880
8	74660410	65536	0	0.9846216220886009	0.9993413713571807	0.0014564244962000236	995
9	74660409	262144	222	0.9705557600146552	0.9985147922913603	0.006931720087857516	218
10	74660408	1048576	21465	0.9436614785175029	0.9971537858968901	0.07851120438233851	107
11	74660407	4194304	493811	0.8926055010656452	0.9953706529867176	0.9776231488048637	70
12	74660406	16777216	5608066	0.8001648289991887	0.993306003164467	5.578247159562147	44
13	74660405	67108864	40177126	0.6493415619698286	0.9909158568215997	25.037656612503476	38
14	74660404	268435456	223323374	0.45266363680539423	0.9881001953197607	65.39414908738563	20
15	74660403	1073741824	1016823080	0.28272035981375565	0.9846937760647562	99.66079650073276	18

Table A.66: *Homo sapiens*, results for chromosome 18, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	56037509	4	0	0.9987513542045561	0.9971686333317393	4.499003522673464E-6	4
2	56037508	16	0	0.9984789116603829	0.9998699485766938	9.922358887975911E-6	16
3	56037507	64	0	0.997918269276326	0.9998806340404804	2.0966729577820377E-5	64
4	56037506	256	0	0.9967448943926948	0.9998166987553236	5.190255492975295E-5	256
5	56037505	1024	0	0.9948523939458047	0.999715063294815	1.2748683388855527E-4	972
6	56037504	4096	0	0.9918715330361609	0.9995517400789568	3.239644244618918E-4	2546
7	56037503	16384	0	0.9865786132547698	0.9992818158200478	9.416276252377555E-4	2054
8	56037502	65536	0	0.977603141553312	0.9989534669476695	0.0029618846167685643	423
9	56037501	262144	56	0.960685952073416	0.9986246742940885	0.01035677930875983	196
10	56037500	1048576	12744	0.9290942850769573	0.9983078837874305	0.09595197209346348	110
11	56037499	4194304	466733	0.8705421703420418	0.9979904900899692	1.3789769470309485	84
12	56037498	16777216	6015193	0.7689355438388773	0.9976282481831596	9.35926300112628	64
13	56037497	67108864	45219514	0.6200584226665227	0.9971842266253294	35.0727248383445	50
14	56037496	268435456	237081429	0.4636731091624794	0.9966336067837286	67.36611976630412	50
15	56037495	1073741824	1037042751	0.3539216733367543	0.995952189464123	88.58327215373511	30

Table A.67: *Homo sapiens*, results for chromosome 19, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	59505253	4	0	0.9959038070134749	0.9965193912217896	6.28068768035944E-5	4
2	59505252	16	0	0.9951235564887617	0.9986837282506296	1.1791848775650811E-4	16
3	59505251	64	0	0.9939190744695792	0.9989075131313213	1.7360706925838326E-4	64
4	59505250	256	0	0.9929463198625331	0.9989579721600491	2.364930167754663E-4	256
5	59505249	1024	0	0.9916238481751417	0.9989349816175059	3.180822265487665E-4	936
6	59505248	4096	0	0.9898788758934338	0.9988007780381802	4.752475432651862E-4	2333
7	59505247	16384	0	0.9863731512617703	0.9984847247676959	9.173357777527682E-4	2326
8	59505246	65536	0	0.9791496198503238	0.9978959019832563	0.002396069528300462	289
9	59505245	262144	157	0.9640441947596384	0.9970053895282636	0.008937707575847198	104
10	59505244	1048576	19011	0.9341753644435102	0.9959420962946117	0.09620825962551251	83
11	59505243	4194304	501832	0.8767368952682035	0.9948235926434525	1.2695334565992715	68
12	59505242	16777216	5787952	0.773127365820981	0.9935939813854224	7.531547350654355	54
13	59505241	67108864	42309006	0.6091634180592597	0.9921233622237082	32.793927755557284	28
14	59505240	268435456	230086135	0.4126969994575268	0.9902759914065061	74.50293629016016	24
15	59505239	1073741824	1027710619	0.26243847201420367	0.9880011580250594	104.09565375119297	16

Table A.68: *Homo sapiens*, results for chromosome 20, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	35449598	4	0	0.9973175436291266	0.9994026539995444	2.4980754458266282E-5	4
2	35449597	16	0	0.9968620517745237	0.9995867102003214	4.686758606061208E-5	16
3	35449596	64	0	0.9955690327190189	0.999601338022544	8.414633772544837E-5	64
4	35449595	256	0	0.9946991495953621	0.9995263128879701	1.5703344963455138E-4	246
5	35449594	1024	0	0.9932563402559702	0.9992765997464071	3.5945640616785393E-4	880
6	35449593	4096	0	0.9905721907724018	0.99865472193082	8.684720105183219E-4	2084
7	35449592	16384	0	0.985106739733422	0.9974025328203413	0.002150949736600664	900
8	35449591	65536	0	0.9745582396141045	0.9950790069199096	0.005559476191614844	129
9	35449590	262144	799	0.9539344742774176	0.9915458105307461	0.02279233203670393	48
10	35449589	1048576	42544	0.9145968941981245	0.9873485097633815	0.27361922126756477	18
11	35449588	4194304	813911	0.8412233168972232	0.9831658650697103	2.656333863577523	10
12	35449587	16777216	7540894	0.7149249721865589	0.9794728395118268	13.724459093691792	8
13	35449586	67108864	49047624	0.531162536002536	0.9757290967693715	47.777910587947204	2
14	35449585	268435456	243396669	0.34311278961375713	0.9712546640275382	87.05676448839911	2
15	35449584	1073741824	1045062167	0.21860499124615962	0.9658343103204001	110.88709237240585	0

Table A.69: *Homo sapiens*, results for chromosome 21, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	35058629	4	0	0.9985429264789561	0.9963931799435148	8.51491035751376E-6	4
2	35058628	16	0	0.9982064899972697	0.9997915615879052	1.6168712452897464E-5	16
3	35058627	64	0	0.9976413508720693	0.9998602556921014	2.4072241995521802E-5	64
4	35058626	256	0	0.9969760937008769	0.9998649397647399	4.168243200306213E-5	254
5	35058625	1024	0	0.9955258085563824	0.9998152430115552	9.66225900314623E-5	914
6	35058624	4096	0	0.9928938739866117	0.9996753061481326	2.8939870548809027E-4	2205
7	35058623	16384	0	0.9872953652515103	0.9993709506993201	9.681324068930729E-4	1032
8	35058622	65536	0	0.976592605379641	0.9988425108410832	0.0034226774313715507	246
9	35058621	262144	371	0.9559830091434571	0.998153160821172	0.01666924475515152	162
10	35058620	1048576	30999	0.9166830297370518	0.9973684468486931	0.2388019946943493 5	94
11	35058619	4194304	713882	0.8432728054690346	0.9964876287484805	2.598049985676052	58
12	35058618	16777216	7384108	0.7173115038362322	0.9954804262936082	15.181266394562272	42
13	35058617	67108864	49555138	0.5401457222342798	0.9942492996268649	48.95184293661661	38
14	35058616	268435456	244823680	0.3696868410321731	0.9927334874750066	84.19261567182077	24
15	35058615	1073741824	1047066658	0.26180321156440434	0.9908423668343397	104.52163538342145	16

Table A.70: *Homo sapiens*, results for chromosome 22, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	152538530	4	0	0.9991222414428669	0.9999612217888257	2.2546602567197917E-6	4
2	152538529	16	0	0.9990783967767252	0.9999648798327465	4.5899761421852336E-6	16
3	152538528	64	0	0.9986692673473287	0.9999659411699638	7.1771323055063775E-6	64
4	152538527	256	0	0.998315632089459	0.999960306832398	1.2079617768301467E-5	256
5	152538526	1024	0	0.9975861573488655	0.9999460757562437	2.6799436921656162E-5	1012
6	152538525	4096	0	0.9962242915355317	0.99990603220432	7.370199452155734E-5	3312
7	152538524	16384	0	0.9935853188142819	0.9998060572812848	2.4391069812575755E-4	8994
8	152538523	65536	0	0.9885523671944824	0.9995745112408111	8.675404733919554E-4	5967
9	152538522	262144	32	0.9787160255820494	0.9991190256953676	0.003335629827027448	1018
10	152538521	1048576	10337	0.9602512994078394	0.9984076216561331	0.02809055422307954	408
11	152538520	4194304	300657	0.9253559559906573	0.9975198887855046	0.32230650370563213	276
12	152538519	16777216	4254988	0.861401846965618	0.9965365972144689	2.4536991692877383	214
13	152538518	67108864	33315598	0.7528857334250487	0.9954562217219465	10.993338484968193	180
14	152538517	268435456	201310351	0.5941176745542898	0.9942111014426582	39.02057724664458	156
15	152538516	1073741824	978906582	0.4274602225709342	0.9927566500989355	73.21518692368691	112

Table A.71: *Homo sapiens*, results for chromosome X, oligonucleotide order 1 to 15.

N order	Total oly	Possible oly	Holes	L1	Sc	Kl	Equivalent pairs
1	25652954	4	0	0.9953160949807184	0.9989006790434252	6.399089641291933E-5	4
2	25652953	16	0	0.9946614333250445	0.9991579537817749	1.2430702619486723E-4	16
3	25652952	64	0	0.9926689138934186	0.9991554624563117	2.2383562555465772E-4	64
4	25652951	256	0	0.9903880454143463	0.9989456639804661	4.1200641677221486E-4	216
5	25652950	1024	0	0.9866340518341945	0.9982703510829066	9.169302343703814E-4	558
6	25652949	4096	0	0.9821815417790758	0.9966718242182045	0.0022840161302602868	1160
7	25652948	16384	0	0.9751053173303903	0.9932732839259476	0.005736969520635962	288
8	25652947	65536	23	0.9628176443041807	0.9860517739243435	0.014490911919967814	33
9	25652946	262144	3400	0.9401696007936087	0.9737361983864715	0.05950516954981247	6
10	25652945	1048576	98537	0.8996320695343166	0.9575165665506614	0.5666268635564358	8
11	25652944	4194304	1255934	0.8299728093586451	0.940790523441896	3.2354835819407333	4
12	25652943	16777216	9473703	0.7224393318146771	0.9254831807034081	14.892109112837238	4
13	25652942	67108864	54618394	0.5914606597559063	0.9112797205271496	40.42275681722711	2
14	25652941	268435456	252463355	0.4866589760604837	0.8961085312975742	62.65675818982261	2
15	25652940	1073741824	1055964898	0.42597417683898997	0.8803447350701034	75.10715409048407	0

Table A.72: *Homo sapiens*, results for chromosome Y, oligonucleotide order 1 to 15.

Bibliography

[1] V. Afreixo, A. Freitas, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, J. M. O. S. Rodrigues, and P. J. S. G. Ferreira. Are inverse symmetries in the human genome statistically significant? *Proceedings of XIIIth Spanish Biometry Conference and 3rd Ibero-American Biometry Meeting*, 2011.

[2] G. Albrecht-Buehler. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proceedings of the National Academy of Sciences*, 103(47):17828–17833, 2006.

[3] G. Albrecht-Buehler. Inversions and inverted transpositions as the basis for an almost universal format of genome sequences. *Genomics*, 90(3):297 – 305, 2007.

[4] P.-F. Baisnee, S. Hampson, and P. Baldi. Why are complementary DNA strands symmetric? *Bioinformatics*, 18(8):1021–1033, 2002.

[5] S. J. Bell and D. R. Forsdyke. Accounting units in DNA. *Journal of Theoretical Biology*, (197):51 – 62, 1999a.

[6] E. Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Cellular and Molecular Life Sciences*, 6:201–209, 1950. 10.1007/BF02173653.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2003.

[8] R. Dahm. Discovering DNA: Friedrich miescher and the early years of nucleic acid research. *Human Genetics*, 122:565–581, 2008. 10.1007/s00439-007-0433-0.

[9] D.Vakatov. *The NCBI C++ Toolkit Book*. National Center for Biotechnology Information (US), 2004.

[10] M. C. E. Bodden and J. Kneis. Arithmetic coding revealed - a guided tour from theory to praxis. technical report. *School of Computer Science, Sable Research Group, McGill University*, May 2007.

[11] D. R. Forsdyke and J. R. Mortimer. Chargaff's legacy. *Gene*, 261(1):127 – 137, 2000.

[12] L. Fredholm. The discovery of the molecular structure of DNA - the double helix. *Nobelprize.org*, Nov 2010.

[13] S. Golomb. Run-length encodings (corresp.). *Information Theory, IEEE Transactions on*, 12(3):399 – 401, jul 1966.

[14] J. J.McEntyre. *The NCBI Handbook*. National Center for Biotechnology Information (US), 2002.

[15] S.-G. Kong, W.-L. Fan, H.-D. Chen, Z.-T. Hsu, N. Zhou, B. Zheng, and H.-C. Lee. Inverse symmetry in complete genomes and whole-genome inverse duplication. *PLoS ONE*, 4(11):e7553, 11 2009.

[16] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[17] Oracle. The java tutorials. `http://docs.oracle.com/javase/tutorial/`.

[18] K. Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, (186):343 – 414, 1895.

[19] A. Pinho, A. Neves, C. Bastos, and P. Ferreira. Dna coding using finite-context models and arithmetic coding. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1693 –1696, april 2009.

[20] D. Qi and A. J. Cuticchia. Compositional symmetries in complete genomes. *BIOINFORMATICS*, 17(6):557–559, 2001.

[21] R.F.Rice. Some practical universal noiseless coding techniques. *Technical Report JPL Publication 79-22*, March 1979.

[22] T. N. Shioiri C. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *Genome Informatics*, 12:433–434, 2001.

[23] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, Apr. 1953.

[24] S.-J. Wei, M. Shi, X.-X. Chen, M. J. Sharkey, C. van Achterberg, G.-Y. Ye, and J.-H. He. New views on strand asymmetry in insect mitochondrial genomes. *PLoS ONE*, 5(9):e12708, 09 2010.

[25] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *Information Theory, IEEE Transactions on*, 23(3):337 – 343, May 1977.